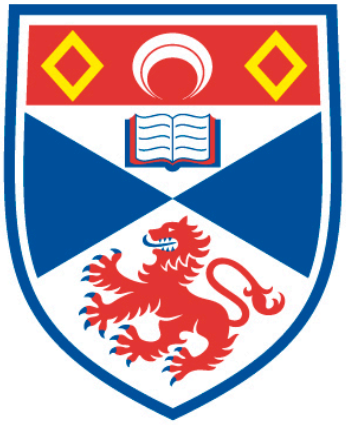
AS THE CROW, FLIES : IDENTIFYING GENOMIC LOCI CONTRIBUTING TO ADAPTATION IN DROSOPHILA AND THE CORVID RADIATION

Ralf Axel W. Wiberg

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

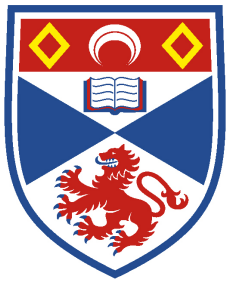


2019

As the crow, flies: Identifying genomic loci contributing to adaptation in *Drosophila* and the corvid radiation.

by

R. Axel W. Wiberg



University of
St Andrews

This thesis is submitted in partial fulfillment for the degree of PhD at the University of St Andrews

December, 2017

Abstract

Identifying loci that contribute to adaptive traits is an important goal of evolutionary biology. I take a comparative genomic approach to identify loci that have responded to divergent selection. First I consider the challenges of identifying consistent genomic responses to selection during experimental evolution. I use population genetic simulations to show that commonly applied statistical tests perform poorly and identify superior methods. These will also be useful in comparing allele frequencies in other contexts. Next I analyse whole genome data from an experimental evolution study of *Drosophila pseudoobscura* evolving under altered mating systems. I find that around 300 SNPs show consistent allele frequency differences between experimental treatments. These are clustered in genomic regions which also show signatures of selective sweeps or background selection. These regions contain genes with mutant phenotypes related to changes already documented in this system. In another chapter I use a novel approach to identify markers potentially influencing female re-mating rate among lines of *D. pseudoobscura*. I use simulations to show that there are more fixed differences between extreme pairs of isofemale lines from different populations than expected by chance. Many of the genes are implicated in female mating behaviour in other studies. I then focus on local adaptation in wild *Drosophila montana* populations. I use Bayesian methods to relate genetic and environmental differentiation among populations. Finally, I take a broader comparative approach using multiple genomes from 14 species of crow to identify potential signatures of selection in the New Caledonian (NC) (*Corvus moneduloides*) and Hawai'ian crow (*Corvus hawaiiensis*) lineages. The NC and, more recently, Hawai'ian crows are of great interest for their unusual tool-using foraging behaviour. I find only modest evidence for greater rates of molecular evolution at coding regions within these lineages. This thesis applies novel techniques to genomic data to identify candidate loci for evolutionary divergence in these different systems and highlight analytical methods that will be generally useful in other systems.

Candidate's declarations

I, R. Axel W. Wiberg, hereby certify that this thesis, which is approximately 60,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

Acknowledgements to other individuals who contributed towards the data and work presented here are made at the start of chapters as appropriate.

I was admitted as a research student in September, 2013 and as a candidate for the degree of PhD in September, 2014; the higher study for which this is a record was carried out in the University of St Andrews between 2013 and 2017.

_____ _____

Date Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

_____ _____

Date Signature of Supervisor

Permission for publication

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis: We request an embargo on the print and electronic copy of this thesis for a period of two years, on the grounds that the publishing of this thesis would preclude furture publishing of the work presented in this thesis. We do not require an embargo on the title or the abstract of the thesis.

_____ _____

Date Signature of Candidate

_____ _____

Date Signature of Supervisor

Acknowledgments

First and foremost I would like to thank my supervisor Professor Mike Ritchie for his support and for his endless patience as deadlines were set, quickly reached, and passed at a dizzying pace, especially during these last few months. He gave me the freedom and courage to explore and develop ideas on my own throughout these last four years (always reigning me in when I got lost) and I have greatly enjoyed working with him.

In the academic sphere there are other people who deserve attention here too. Christian Rütz deserves special mention as a collaborator within St Andrews who also invited me into his lab group to discuss crow behaviour to take my mind off genes once a week. I want to thank Michael Morrissey and Oscar Gaggiotti for discussion and advice on performing population genetic simulations. Thanks to all of the staff for putting up with the endless lab chats on all my problems and for the subsequent illuminating discussions in the pub. I also want to thank my other collaborators I have met during the projects presented here. So thanks to Rhonda Snook, Tom Price, Nina Wedell, Maaria Kankare, and Anneli Hoikkala for introducing me to different systems of *Drosophila* and the questions that excited them, and also, again, for their patience with me.

Thanks also go to the technicians and support staff within the Harold Mitchell Building and Dyers Brae, without whom I would still be floundering waist deep in *Drosophila* food and admin.

I also want to thank my friends, past and present, within the Harold Mitchell Building and Dyers Brae for keeping me sane. They've always been there for a beer and a laugh when the going got tough and helped me keep a reasonable perspective on life. Thanks especially to Will for being an amazing flatmate and reminding me that music is wonderful thing.

Finally, I want to thank my partner Deirdre, my parents, and family who have been loving, understanding, and supportive throughout my studies.

During my PhD studies I was also a co-author on the following papers through collaborations with other groups. These studies do not form a part of this thesis.

DrosEU Consortium. Population Genomics of European *Drosophila melanogaster*. **[*in prep*]**

D. J. Parker, **R. A. W. Wiberg**, P. Veltsos, U. Trivedi, V. I. Tyukmaeva, K. Gharbi, R. K. Butlin, A. Hoikkala, M. Kankare1, and M. G. Ritchie The Genome of *Drosophila montana* a cryophilic fly. **[*in prep*]**

Christopher B. Cunningham, Lexiang Ji, **R. Axel W. Wiberg**, Jennifer Shelton, Elizabeth C. McKinney, Darren J. Parker, Richard B. Meagher, Kyle M. Benowitz, Eileen M. Roy-Zokan, Michael G. Ritchie, Susan J. Brown, Robert J. Schmitz, Allen J. Moore. 2015. The genome and methylome of a beetle with complex social behavior, Nicrophorus vespilloides (Coleoptera: Silphidae). *Genome Biology and Evolution* **7**:3383-3396

Table of Contents

"...endless forms most
beautiful...have been, and are
being evolved."
- C. Darwin 1859

Chapter 1 Introduction

1.1 Genes, Genomes, and Adaptation

With the publication of "*On the Origin of Species"* in 1859 Charles Darwin introduced natural selection (Darwin 1859). A force which, he argued, produces the remarkable appearance of an organism fit for its environment. A character that results in an individual leaving on average more offspring in the next generation than other individuals (i.e. it is adaptive), if it is heritable, will spread and become more common in the population. Since Darwin we have learned that, by and large, genes underlie the heritable component of a phenotype. To the extent that a phenotypic trait has some genetic component the trait can be passed between parents and offspring. If the trait is adaptive the alleles producing the phenotype will become more frequent in the population. This combination of Darwinian natural selection with genetical inheritance of traits through generations is typically presented as the "modern synthesis" of evolutionary biology (Huxley 1942) and forms the foundation from which we can understand biological diversity. Given the understanding that evolution acts on allelic variants to produce phenotypic change, efforts to understand the genetic basis of variation in traits important in adaptation, sexual selection and speciation have been the focus of much research throughout the last century.

This research program takes aim at several questions in various fields. In the study of animal behaviour it has long been recognised that the genetic basis of traits represents an important component of understanding how they evolve (i.e. Tinbergen's "four questions"; Tinbergen 1963). Continued calls have been made to incorporate modern genetic methods and tools to questions of ontogeny and mechanism in animal

behaviour (Fitzpatrick et al., 2005; Zuk and Belanger, 2014) but also the evolution of it. An emerging view is that gene expression differences are at the heart of behavioural differences. For example, the genetic control and evolutionary origins of behaviour have been investigated by studying gene expression during different parenting phases in burying beetles (Parker et al., 2015; Cunningham et al., 2017), social behaviour in honey bees and other social insects (Ritschoff & Robinson 2012; Zayed & Robinson 2012), and others (see e.g. Ritchie & Butlin 2014; Rittschoff & Robinson 2014). However, the regulatory loci which are the targets of selection to produce these differences in gene expression (and behaviour) between populations and species are less well understood.

More recently, calls have also been made to bring genomics into the investigations of sexually selected traits which are some of the most diverse and striking traits in nature (Wilkinson et al., 2015). Here too, there is emerging evidence that gene expression (as opposed to coding sequence changes) plays a large role. Expression differences between males and females are potentially important in resolving intra-genomic sexual conflicts and producing sexual dimorphic morphologies and behaviours (Wilkinson et al., 2015). However, the genomic targets of selection that produce these regulatory differences between populations and species remain relatively unknown.

In the study of speciation comparative genomics can bear on fundamental questions. For example, the role of introgressions and hybridisation during speciation, the loci influencing reproductive isolation and ecological adaptation, as well as how isolation can build up between populations in the face of continued gene flow are of great interest (Barrett and Hoekstra 2011; Radwan and Babik 2012; Lee et al., 2014; Seehausen et al., 2014). Other questions concern how genes that underlie these important traits are distributed throughout the genome. Do differences in genomic organisation (e.g. inversions and the recombination landscape) contribute to or constrain adaptive evolution (Hoffmann & Rieseberg 2008; Radwan and Babik 2012; Lee et al., 2014)?

The past few decades have seen the rise and rapid development of modern genetics and genomics. Accompanying this expansion is the cottage industry of bioinformatic pipelines and software tailored to the analytical needs of researchers. These developments have allowed researchers to sequence the entire genomes, of first

model and more recently non-model organisms. Comparison of these genomes allow the identification of regions and potentially individual nucleotides that contribute to variation in adaptive phenotypes (see below for an overview of some successful case studies).

However, these questions are complex and there are many difficulties to overcome that vary depending on the nature of the trait and system in question (Mackay 2009; Rockman 2012; Travisano and Shaw 2013; Lee et al., 2014). Several authors have argued that many of the loci discovered to underlie variation in trait within populations to date are large-effect loci controlling essentially Mendelian or discrete traits. The same critics point out that our understanding of complex and quantitative or continuous traits remains poor because of their likely polygenic underpinnings (Mackay 2009; Rockman 2012; Travisano & Shaw 2013; Lee 2014). A similar problem is thought to apply in attempts to identify loci of speciation or barrier loci (Noor & Bennett 2009; Wolf & Ellegren 2017; Ravinet et al., *in press*). Similar patterns of genetic diversity within and between species can arise by selection or by demographic, essentially neutral processes which makes inference difficult (Noor & Bennett 2009; Wolf & Ellegren 2017; Ravinet et al., *in press*).

While, understanding these difficulties and the limitations of different methods is important, this should not cause us to "throw up our hands" (Rockman 2012) but rather to roll up our sleeves. Understanding the genetic basis of traits and adaptation is necessarily going to be difficult because of the many complexities involved and the hypotheses or questions driving the research should be always in view (Zuk & Belanger 2014). In many questions, such as whether parallel phenotypic evolution is a result of parallel genetic changes, a genetic perspective is essential (Rausher & Delph 2015). Additionally, novel innovative experimental designs and technologies (e.g. pooled sequencing, CRISPR/*Cas9*) along with falling costs will help to increase sample sizes and power. Comparative genomics (whether it be comparing different experimental evolution treatment lines, different populations, or different species) is a useful starting point and a valuable complement to other experiments to uncover the loci underlying population and species differences. It is in this context that I have done the work presented in this thesis.

1.2 The Tools of the Trade: Methods and Approaches in Comparative Genomics

1.2.1 The Next (Sequencing) Generation: "Engage!"

At the turn of the millennium the human genome was finally complete and was hailed as another milestone for the field of biology. Since then, the sequencing technology as well as analytical methods have improved dramatically (see e.g. Schuster 2008; Ansorge 2009; Metzker 2010; Ekblom and Wolf 2014 for reviews). The new technologies have collectively become known as "next-generation sequencing" (NGS) and are characterised, in general, by higher throughput, being able to sequence 10s or 100s of individual genomes in a short period of time, as well as lower costs. Initially very expensive NGS technologies were still only applied to model organisms and economically important domestic animals (e.g. chicken, sheep, mouse, *Drosophila*, and human populations). However, declining costs quickly allowed access to these technologies for smaller research groups working on interesting traits in non-model species (Stapley et al., 2010; Ekblom and Wolf 2014). Many thousands of species now have genomes sequenced at various levels of completeness as well as accompanying population genomic data about allele frequencies at identified genomic markers. While NGS sequencing technologies are cheaper and faster than traditional sequencing (Sanger) they generally produce much shorter reads and as such genome assembly is more difficult, especially in long repeat-rich regions. These technologies are still advancing and longer reads are slowly becoming possible through PacBio (Pacific Biosciences) and Illumina mate-pair or synthetic long reads (Illumina) which can exceed the length of traditional Sanger sequencing (Schuster 2008).

Researchers are also getting creative with applying these technologies to their needs. For example, a relatively novel application of short-read whole genome sequencing is to sequence pools (known as "pool-seq") of individuals rather than producing individual genomes (e.g. Schlötterer et al., 2014). This is useful if researchers require accurate data only on the frequencies of alleles at genomic markers and are not interested in preserving haplotype information for individuals. This can allow very accurate estimation of even low frequency allelic variants for a fraction of the cost it

would take to produce individual whole-genome sequences (Schlötterer et al., 2014; Wolf and Ellegren 2017).

Data on gene expression is also now readily available even for non-model species via RNA-sequencing (RNA-seq) (Marguerat and Bähler 2010; Wolf 2013). mRNA from whole-organisms, different tissues and under different environmental conditions can be extracted. This mRNA can then be reverse-transcribed to cDNA after which standard NGS can be performed (Marguerat and Bähler 2010). This method has become very popular and allows the characterisation of the level of expression of all genes in a highly targeted manner. Genes that show evidence of being differentially expressed between two phenotypic classes, environments, experimental treatments can then be used as candidates to explore the genetic basis of different traits. For example, different caste types in honey bees can be compared to uncover the genes involved in producing differences in behaviour (Zayed & Robinson 2012).

Finally it is now becoming possible to follow up on many interesting loci discovered by functional validation of particular variants or alleles via functional genomics. Technologies such as genome editing by CRISPR-cas9 (Jinek et al., 2012; Bassett et al., 2013) or the manipulation of transcription by RNA interference (RNAi, Kamath & Ahringer 2003; Boutros et al., 2004).

1.2.2 Genome Wide Association (GWA), Quantitative Trait Locus (QTL) Mapping Studies, and Comparative Genomics

QTL mapping started in the 1980s, and progressed to GWA studies (GWAS) when more genomic markers and larger sample sizes become available. QTL mapping relies on crossing experiments to map associations between regions and a phenotype via recombination (Stinchcombe & Hoekstra 2008). GWAS takes advantage of historical recombination in natural cohorts or populations (Stinchcombe & Hoekstra 2008; Visscher et al., 2012) or diversity panels of inbred lines (e.g. Ivanov et al., 2015). Broadly, these methods are related in that the aim is to map genomic markers linked to loci that underlie variation in interesting traits (e.g. adaptations, or complex diseases). However, these methods typically require either extensive breeding designs (QTL mapping, or inbred lines) or enormous sample sizes. Even then, the results suggest that only a very small amount of the variation in a trait is explained by QTLs or GWAS loci

(Mackay 2009; Rockman 2012; Travisano & Shaw 2013). Additionally, the effect sizes of moderate effect loci may be overestimated due to many linked low effect loci (Mackay 2009; Rockman 2012; Travisano & Shaw 2013).

Thus, although QTLs or GWAS can in principle identify variants that underlie variation in known adaptive traits (often referred to as "forward genetics"; Ellegren 2014; Pardo-Diaz et al., 2015) it has some drawbacks. By contrast, a comparative genomic approach (or "reverse genetics"; Ellegren 2008; 2014; Pardo-Diaz et al., 2015) aims to uncover the loci which contribute to differences between species, rather than a complete account of all loci that underlie variation in a trait (Ellegren 2008; Ellegren 2014; Pardo-Diaz et al., 2015). For example, differences in coat or plumage colour (e.g. Hoekstra et al., 2006; Steiner et al., 2007; Poelstra et al., 2015; Vijay et al., 2016), or the differences in adaptive skeletal morphologies (e.g. Jones et al., 2012; Lamichhaney et al., 2015) between closely related species or populations can be studied to find loci that contribute to these diferences. In the study of evolution and the genomic targets of selection, a comparative approach can be a fruitful approach to identify loci that underlie differences between populations and species.

1.2.3 Evolution in the Lab: Experimental Evolution Studies

Experimental evolution involves the investigation of how successive generations evolve in response to a novel environment. This environment can be a novel food source (Lenski 2011), a new temperature regime (Orozco-terWengel et al., 2012), an altered social environment (Crudgington et al., 2005; Janicke et al., 2016), exposure to pathogens (Zbinden et al., 2008; Martins et al., 2014) or almost any other environmental variable that an experimenter can control. Experimental evolution and selection experiments have been highly productive and are growing in popularity (Kawecki et al., 2012). Experimental evolution is not a new experimental paradigm; forays by the Reverend William Dallinger, a contemporary of Charles Darwin, date back to the latter half of the 19th century (Lenski 2011).

The combination of experimental evolution with NGS technologies in what has come to be called "Evolve & Re-sequence" (E&R) studies is an emerging trend (Kawecki et al., 2012; Kofler and Schlötterer 2013). Such studies are rapidly becoming popular approaches for understanding how organisms adapt to novel or different

environments. These types of experiments are also useful for understanding other processes of evolution, such as mutation and genetic drift. For example, the rate at which deleterious and harmful mutations arise can be studied in mutation accumulation experiments (Ness et al., 2015; Morgan et al., 2014). Simple environmental variables allow a researcher to set up multiple replicates of the same treatments in order to test whether responses to selection are consistent across these replicates. The use of multiple replicates allows experimenters to distinguish between changes in just a few lines due to random drift compared to changes across all lines which are more likely due to selection (Kawecki et al., 2012; Kofler & Schlötterer 2014**)**. Similarly, the genomes of ancestral and descendant populations or descendant populations of different treatments can be sequenced to find alleles that have changed in frequencies across the replicated treatments. Such markers are strong candidates for loci that underlie variation in the new phenotypes. In systems that allow for it, generations can be literally frozen in time in a resting state and competition experiments set up to study the fitness consequences of particular mutations (Lenski 2011; Ness et al., 2015).

There are some obvious limitations to these sorts of studies. First, they are restricted to organisms that will readily breed in captivity and/or survive in laboratory environments. Second, the organisms need to have generation times that make it feasible to run the study for several generations. Finally, but most importantly, there must be some genetic variation underlying the phenotypes in question. Novel, adaptive mutations are unlikely to arise in the short timescales of most experimental evolution studies (Kawecki et al., 2012; Kofler & Schlötterer 2014). Nevertheless, experimental evolution studies are a valuable tool for biologists to test predictions from theory about what sorts of responses that should occur in response to different sources of selection (e.g. sexual conflict or sexual selection), study specific adaptations and their genetic basis as well as the relative importance of adaptation from standing genetic variation or novel mutations (Kawecki et al., 2012). However, there are still challenges to overcome as well as novel analytical methods needed to study experimental evolution in the genomic era. The increasing availability of datasets and the rapid development of sequencing technologies and follow-up functional genomic tools runs the risk of generating an abundance of data before the analytical methods have had time to be developed and optimised.

1.2.4 Comparing and Contrasting Populations or Species

The comparative approach has a long tradition in evolutionary biology. The comparison of different closely related species that differ in a trait of interest has been used to study adaptation to different environments (Harvey and Purvis 1991; Harvey and Pagel 1991; Arnqvist and Rowe 2005). Similarly, different populations of the same species can show local adaptation to particular parts of the range. Such systems allow the study of the forces that shape different adaptations. In the genomic era such systems also have the potential of identifying genomic regions that underlie adaptive differences between populations in these traits (Ellegren 2008; Ellegren 2014; Pardo-Diaz et al., 2015). By comparing closely related species or populations that vary in the trait of interest and (ideally) not many other characteristics researchers can identify regions that show higher genetic differentiation than the genome-wide background. However, demographic processes can lead to very similar signals which can make inference difficult.

An enormous amount of population genetic theory has been brought to bear in the aim of comparing different populations. Methods have been developed to identify population structure and admixture between populations (e.g. STRUCTURE; Pritchard et al., 2000 and fineSTRUCTURE; Lawson et al., 2012), adaptive introgressions (*D* statistic or ABBA-BABA; Green et al., 2010), selection and local adaptation (Foll & Gaggiotti 2008; de Villemereuil & Gaggiotti) and various statistics to detect demographic effects and historical migration (e.g. ABC analysis Csilléry et al., 2010). Such statistics (F_{ST}, Tajima's D, π, etc.) are well founded in population genetic theory but our expectations from theory are based on certain assumptions about the behaviour of these statistics during evolution. These are necessary to distinguish between different hypotheses of which forces have produced the observed patterns.

For example, in population genomic studies, a common and popular approach has been to identify outlier loci by measuring genetic differentiation (F_{ST}) at many loci throughout the genome and focus on the tails of the distribution to identify outliers. However, many recent reviews have demonstrated that variation in F_{ST} throughout the genome could arise from a number of causes and these are not always easy to distinguish on the basis of a single summary statistic. F_{ST} peaks can also be produced by

variation in recombination throughout the genome, ancestral low diversity regions, and other processes. This makes their use to infer differentiation specifically due to barriers to gene flow or selection problematic (Noor and Bennet 2009; Wolf and Ellegren 2017; Ravinet et al., *in press*). For example, variation in the effective population size throughout the genome (e.g. due to demographic expansion), and nucleotide diversity can lead to heterogeneity in levels of F_{ST}. This can result in the identification of many outlier loci which are not the product of selection (Wolf and Ellegren 2017). Similarly, regions of reduced gene flow (e.g. due to loci that result in reproductive incompatibility) will show higher F_{ST} (Wolf and Ellegren 2017). Recent work has also shown that if the recombination landscape is very similar across species, then within species diversity can be highly correlated across pairwise comparison (Dutoit et al., 2017). Since F_{ST} is a relative measure of diversity within and between species/populations (Cruickshank & Hahn 2014; Wolf and Ellegren 2016), this means that interpreting the landscape of a single measure of differentiation between two species or populations can be misleading. A potential solution may be to compare several population genetic summary statistics to see if there are any population specific "peaks" (Wolf and Ellegren 2017) and also, if possible, to compare several independent pairwise contrasts (e.g. Lamichhaney et al., 2015; Vijay et al., 2016).

In the case of some traits, more information can be obtained from careful sampling of populations. This is true in the study of clinal phenotypic variation where a transect of populations can be sampled. The continuous variation in environmental variables or phenotypes can then be related to the variation in allele frequencies to uncover relationships (by e.g. Bayesian methods; de Villemereuil & Gaggiotti 2015). Such relationships can identify variants underlying local adaptation in respone to environmental variation. Leveraging information from independent clines in different regions of the world can provide even more power and shed light on the repeatability of adaptive evolution (Adrion et al., 2015).

Comparing more distantly related groups, i.e. closely related species, requires different measures of differentiation. The substitution rate or divergence (d_{XY}) between two homologous regions in a pairwise comparison of species can identify highly conserved or rapidly diverging regions. Contrasting d_{XY} with the within species diversity can identify regions where divergence is higher or lower than expected from the

mutation rate and the neutral substitution rate (e.g. the HKA test: Hudson et al., 1987 contrasting π and d_{XY}; Halligan et al., 2013). This method applies both to coding and non-coding genomic regions. At coding regions more sophisticated analyses can compare the synonymous and non-synonymous substitution rates (*dS* and *dN* respectively) across the coding region of a gene in different lineages on a phylogeny (Yang 1998; Zhang et al., 2005; Yang 2007). Such an approach can identify amino acid changes that evolve in response to selection. These tests are highly dependent on accurate alignments and annotation of coding regions in the species under investigation (Schneider et al., 2009) which limits this sort of analysis to relatively well sequenced and annotated species.

1.3 The Genomics of Adaptation: Case Studies

A system which has produced exciting results recently is the *Heliconius* species complex which comprises multiple species of butterflies exhibiting a striking case of Mullerian mimicry (Sheppard et al., 1985; Mallet 1989 Nadeau et al., 2012;). The species *H. melpomene* occurs in a number of different colour pattern morphs that all resemble the geographically co-occurring *H. erato* morphs (Sheppard et al., 1985; Mallet 1989; Joron et al., 2006). Similarly the species *H. numata* consists of several morphs which mimic separate co-occurring species of another family of butterflies (Joron et al., 2006). In this system the genetic basis of various aspects of the colour pattern has been uncovered first by traditional genetic crosses (Sheppard et al., 1985; Mallet 1989) and more recently by modern genomic methods (Joron et al., 2006; Ferguson et al., 2010; Joron et al., 2011;). A handful of loci control much of the variation in colour patterns seen in *H. melpomene* and *H. erato*, while a single supergene controls different patterns in *H. numata* (Sheppard et al., 1985; Mallet 1989; Joron et al., 2006; Joron et al., 2011). The use of a variety of molecular markers demonstrated that the loci controlling colour patterns in *H. erato, H. melpomene* and *H. numata* are indeed homologous loci (Joron et al., 2006). Further characterisation of the *H. melpomene* transcriptome and the annotation of some of the colour pattern determining loci identified several genes lying within these regions though very few of them have been implicated in colour pattern formation in other species (Ferguson et al.,

2010). Further work shows that the supergene in *H. numata* arises as a result of chromosomal rearrangements that capture different variants and prevent recombination (Joron et al., 2011). Meanwhile, in *H. melpomene* and closely related, co-occurring *H. timareta* mimcry is facilitated by exchange by gene flow of colour pattern loci (The *Heliconius* Genome Consortium 2012). Further work identified spefic genes that control patterning (*WntA*; Nadeau et al., 2014) and pigment deposition (*optix, cinnaba*r; Nadeau et al., 2014; Jiggins et al., 2017). In another butterfly mimicry ring, where only males exhibit the mimetic colour patterns, a sex-specific transcription factor *doublesex* has been co-opted to regulate a suite of genes in a sex specific manner (thus acting as a "supergene") to produce the mimetic colour patterns (Kunte et al., 2014). These studies highlight how regulatory modules are physically shuffled, and co-opted to produce a diversity of wing patterns in butterflies (Jiggins et al., 2016). Although much remains unknown in this system it is clear that comparative genomic approaches have provided great insights into how intricate colour patterns can evolve and be maintained in populations.

A further recent example is the determination of the genetic basis of different male morphs in the ruff (*Philomachus pugnax*) (Lamichhaney et al., 2016; Küpper et al., 2016). This species is a classic example of sexual selection via male-male competition for females that has produced lekking behaviour as well as different mating strategies among male. The genetic basis of this trait has been speculated on previously (Fitzpatrick et al., 2005) and the system was recently mentioned in a review calling for bringing research on sexually selected traits into the genomic era (Wilkinson et al., 2015). Three distinct male morphs exist in this species. "Independents" are males that compete with other males by displaying and defending territory in leks. Meanwhile "satellite" males don't defend territories and are submissive to independent males but can obtain matings from females by their proximity to the independents. Finally, the "faeder" male morph obtains matings by mimicking the appearance of females which allows them to avoid male-male competitive interactions (Lamichhaney et al., 2016; Küpper et al., 2016). Lamichhaney et al., (2016) produced high quality and deep coverage genome sequences of a total of 16 independent males, 8 satellite males and one faeder male and identified a ~2,000 bp inversion with extraordinary levels of differentiation in all parwise comparisons. Interestingly the inversion contains regions

that are highly differentiated between satellites and independents as well as faeders and independents but also smaller regions that are differentiated between faeders and satellites indicating a complex sequence of inversion and recombination events producing the satellite morph (Lamichhaney et al., 2016). Many genes occur within this inverted regions some of which have obvious associations to variation in plumage colouration in birds (MC1R) and others that are involved in the metabolism of sex hormones (Lamichhaney et al., 2016). A completely independent study using standard genome-wide assocation methods in a pedigreed captive population with multiple individuals of each morph found evidence for an inversion in the same region of the genome and even similar breakpoints disrupting the coding sequence of the gene CENP (Küpper et al., 2016). These studies highlight another example where the genetic basis of a complex and multifaceted phenotypic trait, though discrete, involves complex genomic re-arrangements that result in combinations of alleleic variants ('supergenes') that together act to produce the phenotypes.

Another example of the insights gained from comparative genomics comes from natural populations of the cosmopolitan fruitfly *Drosophila melanogaster*. *D. melanogaster* shows adaptive clinal variation in various traits including body size (Kolaczkowski et al., 2011; Chen et al., 2012; Flatt 2016). Research has identified several F_{ST} outlier regions between northern and southern populations in Australia. Genes in these regions are important in signalling pathways that influence metabolism and body size (Kolaczkowski et al., 2011). Meanwhile, Chen et al., (2012) collected flies from several populations of an Australian cline. Larvae from multiple northern and southern populations were reared under identical conditions in the lab and gene expression quantified by microarrays (Chen et al., 2012). Northern and southern populations showed differential expression at several genes associated with metabolism and experimental knockdown of some resulted in smaller body sizes in the predicted direction (Chen et al., 2012). Meanwhile, pool-seq studies of North American populations of *D. melanogaster* find independent support for selection on genes within the same signalling pathways involved in the determination of body size as in Kolaczkowski et al., (2011) (Fabian et al., 2016). However, it has since been discovered that subtle patterns of demographic history and migration contribute substantially to this clinal variation. This makes distinguishing between true adaptive differentiation and

demographic history difficult especially in young species or population clines (Bergland et al., 2016; Flatt 2016). Nevertheless, these studies show that candidate loci can be uncovered in studies of clinal populations that, when followed up with functional genomic studies, produce the expected phenotypic changes. Additionally, convergent clinal patterns have arisen due to similar selection pressures acting on different populations of *D. melanogaster*.

1.4 Thesis Outline and Aims

The studies presented in this thesis are primarily based on comparative genomic methods. I compare experimental evolution treatment lines, different populations and species in an effort to identify loci that underlie observed phenotypic variation. In chapter two of this thesis I address the challenge of identifying consistent allele frequencies differences between two treatments across multiple replicates of an experimental evolution study. Similar problems also arise in population genomic studies where allele frequency differences between different populations are related to some environmental or phenotypic differentiation. I perform population genetic simulations that simulate an experimental evolution study under the neutral model (the null hypothesis). I assess the behaviour of different statistical approaches in identifying consistent allele frequency differences.

In chapter three I apply the methods from chapter two to data from an ongoing, long-term experimental evolution study in *Drosophila pseudoobscura* (Crudgington et al., 2005). The aim was to identify the genomic targets of selection under different mating systems. Previous studies have identified changes in traits like the courtship song (Snook et al., 2005), higher male courtship frequencies (Crudgington et al., 2010), but also in gene expression patterns (Immonen et al., 2014). Thus, we might expect that genomic changes might be enriched for regulatory elements (such as transcription factor binding sites). Additionally, this experimental evolution study allows the testing of predictions from theory about the effects of mating systems and sexual selection on the molecular evolution of the X chromosomes compared to the autosomes (Ellegren 2009).

In chapter four I focus on another aspect of mating system evolution, namely the female re-mating rate. Wild populations of *D. pseudoobscura* show variation in the

female re-mating rate (Price et al., 2011; Herrera et al., 2014). This seems to have direct implications for the spread of sex ratio distorting selfish genetic elements in the population (Price et al., 2010; Price et al., 2014). I use a novel method in an attempt to identify loci associated with the female re-mating rate. This method relies on isofemale lines from different natural populations of *D. pseudoobscura*. These isofemale lines show variation in female re-mating rate that has persisted over several generations of lab maintenance indicating a genetic component to this trait (Price et al., 2011). I perform whole-genome sequencing of pairs of isofemale lines from the extremes of the distribution of re-mating rates. I compare pairs of lines from the same populations to identify SNPs which are consistently fixed for the same alleles in high and low re-mating lines across all populations. Within a single pairwise comparison, many fixations might arise due to chance. I use population genomic simulations under realistic assumptions about the population sizes and mutation rates to estimate how many consistently fixed differences are expected by chance.

In chapter five I describe a population genomic clinal study of *Drosophila montana* populations from North America and Finland. This species has been of interest because of its extreme cold-tolerance and the variation in this tolerance across populations (e.g. Vesala et al., 2012). Another trait that is thought to be related to its cold-tolerance are diapausing behaviour. Because the species occurs throughout the northern hemisphere where there is substantial latitudinal environmental variation in the day length, coldest temperatures, and other environmental variables, there is opportunity for populations occurring in different parts of the clinal range to show evidence of local adaptation. Such local adaptation would allow for the identification of variants that underlie the traits contributing to population divergence in this species. I use cutting-edge Bayesian methods (de Villemereuil & Gaggiotti 2015) to relate genetic differentiation to composite measures of climatic differentiation across populations from different parts of the species range.

In the final chapter, I present a comparative genomic study of crows (genus *Corvus*). The New Caledonian (NC) crow (*Corvus moneduloides*) and more recently the Hawai'ian crow (*Corvus hawaiiensis*) have attracted much attention due to their tool-using foraging behaviour (Emery & Clayton 2004). The NC crow is known to manufacture and use different stick tools in the wild in order to aid in foraging for beetle

larvae (Rutz & St Clair 2012). Meanwhile, the Hawai'ian crow, which is extinct in the wild, has recently been shown to have the capacity for tool manufacture and use in captivity (Rutz et al., 2016). Both of these species show similar skull and beak morphological features that are known in the NC crow to aid in the use of tools (Rutz & St Clair 2012; Troscianko et al., 2012; Matsui et al., 2012; Rutz et al., 2016). This raises the possibility that these are adaptations to a tool-using lifestyle in the wild (Rutz & St Clair 2012; Troscianko et al., 2012; Matsui et al., 2012; Rutz et al., 2016). The comparative genomic study aims to identify signatures of selection throughout the genome. Using an "ecological control" species which is also a tropical island endemic which has presumably experienced a similar demographic history (population contractions, adaptation to island habitats, etc.) I try to identify genomic loci that show evidence of selection within the NC crow.

Statement

A version of the following chapter:

"Identifying Consistent Allele Frequency Differences in Studies of Stratified Populations"

has been published in the journal *Methods in Ecology and Evolution* under the same title with the following co-authors:

Michael M. Morrissey, Oscar E. Gaggiotti, and Michael G. Ritchie.

My contributions to the submitted manuscript and to this chapter were; the conception of the project, development, running and analysis of all simulations. I led the writing of the manuscript and submission process. Throughout I receieved advice from my co-authors on the practical design of the simulations, the presentation and evaluation of the results, and on the text of earlier drafts of the manuscript.

"To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of."

- R. A. Fisher ca. 1938

Chapter 2 Identifying Consistent Allele Frequency Differences in Studies of Stratified Populations

Abstract

Experimental evolution studies are a powerful tool to study adaptation in the laboratory and population genomic technologies are rapidly starting to be applied in these contexts. Several approaches have been proposed to analyse the emerging data in stratified populations such as replicated experimental evolution lines. A commonly recommended and employed statistical method is the Cochrane-Mantel-Haenzel (CMH) test. However, a careful reading of the original literature gives good *a priori* reasons to think that the CMH-test is ill suited for the analysis of allele frequency differences in pool-seq population genomic studies. There are other alternatives that would seem more suitable (e.g. Generalised Linear Models (GLMs) with quasibinomial error distribution). I therefore set out to test, by simulation, these other approaches and compare their performance to the CMH-test.

The simulations in this chapter consider a simple population genetic model that reflects a hypothetical experimental evolution scenario where allele frequencies are only changing due to neutral drift. This is the "Null hypothesis" in these types of studies. A good statistical test should perform well under the null hypothesis and produce well-behaved p-value distributions which do not show an inflation of false positives.

Additionally, I assess the power of different tests by comparing the recovery rate of a small number of simulated "True Positives." The results show that the CMH-test indeed does produce high false positive rates (FPRs). Meanwhile a Generalised Linear Model with quasibinomial error structure (QBGLMs) performs very well under the null hypothesis and does not suffer any loss of power in recovering true positives. Therefore I conclude that QBGLMs are prefereable to other approaches in the analysis of these kinds of data. QBGLMs should also be useful in other types of analysis of population genomic data, for example in relating allele frequency differences among populations to some other variable (e.g. latitude or altitude).

Author Contributions

In this chapter I designed and performed all of the simulations and analyses. I am grateful to Michael Morrissey, Oscar Gaggiotti and David Shuker for invaluable advice and discussion throughout the work presented in this chapter.

2.1 Introduction

With the increasing application of pooled genome sequencing (pool-seq) approaches to population genomics (Boitard et al., 2012; Ferretti et al., 2013; Schlötterer et al., 2014, 2015) researchers are interested in accurately quantifying allele frequencies within and between populations and using these differences to infer the action of selection. Such data can provide us with insight into the evolutionary and demographic history of populations and help to identify regions under selection as well as alleles that consistently differ in frequency between population substrata with different characteristics, across populations.

In particular, several tests of frequency differences have been used to compare allele frequencies at markers throughout the genome. The aim is to determine whether the frequencies of an allele at a particular marker (typically single nucleotide polymorphisms; SNPs) are consistently different between subsets of a population or whether such differences are consistent across replicated experimental evolution lines. This consistency is important because it provides a criterion to identify alleles that underlie the same trait in many populations or to distinguish consistent responses to

selection from idiosyncratic responses or effects of drift in experimental evolution studies. A hypothetical example is where replicate lines of a large mass selection treatment are set up from three separate but identical base line populations and allowed to evolve for several generations. Pooled whole genome sequencing (Pool-seq; Schlötterer et al., 2014) can then be applied to determine the allele frequencies at different SNP markers throughout the genome in the base populations and the last generation. The goal is to find SNP alleles that consistently occur at different frequencies at the end of the experiment across the replicates. Markers that show such a consistent difference are more likely to be functionally important in producing the phenotype under study.

Many of the statistical tests applicable to this kind of scenario are implemented in popular population genomic software tools (e.g. PoPoolation2) which make them routine for researchers to apply. However, here I find that there are serious consequences of the misapplication of these tests that arise from two main sources. First, heterogeneity in allele frequency differences (e.g. arising from genetic drift) is often confused for significant main effects. Second, very little attention has been paid to pseudoreplication of allele counts that is inherent in pool-seq experimental designs. I show that these violations of statistical assumptions produce high false discovery rates (FDRs). These problems are highlighted by simulation and I present alternative tests and filters for the analysis which improve inference.

2.1.1 The CMH-test

Perhaps the most widely used statistical method to compare allele frequencies is the Cochrane-Mantel-Haenszel test (Cochran, 1954; Mantel & Haenszel, 1959)**,** an extension of Chi-squared tests for multiple biological replicates. The CMH-test considers 2x2x*k* contingency tables. In the context of population genomics the rows and columns of each 2x2 table represent counts of different alleles (X) in different treatment lines or strata (Y) while *k* represents the number of biological replicates (e.g. different studies, populations, Z) (Agresti, 1996). In the CMH-test the null hypothesis is that "*X* and *Y* are conditionally independent given Z" (Agresti, 1996). A 2x2 table can be summarised by the conditional odds (O_{XYk}) which measures the magnitude of the association between the factors *X* and *Y*.

If;

$$O_{XY1} = O_{XY2} = \ldots = O_{XYk}$$

Then the odds ratios are homogenous, the association between *X* and *Y* is the same at each level (*k*) of Z, and I are justified in describing the association with a single common odds ratio which can be tested for differences to 1 (Agresti, 1996). However, if the association between *X* and *Y* for the 2x2 tables is different across the *k* tables the test can give misleading results (Landis et al., 1978; Agresti, 1996; see also below):

> "The CMH statistic takes larger values when ($n_{11}k - mu_{11}k$) is consistently positive or consistently negative for all tables rather than positive for some and negative for others. **This test is innapropriate when the association varies dramatically among the partial tables. It works best when the X-Y association is similar in each partial table.**"
> (Agresti 1996, pp 61, emphasis added)

This assumption of homogeneity can be tested by, for example, the Woolf-test (Woolf, 1955). Another assumption of the CMH-test is that data contributing to each count within a cell of the contingency table are independent. The first assumption is frequently violated in real data. Furthermore, it is in fact the pattern of consistency that is of interest. The second assumption is violated automatically in the design of pool-seq experiments because allele counts obtained from reads directly are not independent draws from the treatment line or study population. Note also that this test assumes a pairing between the two treatment lines nested within replicates. Such a pairing may sometimes be biologically meaningful (e.g. if any two treatment and control lines were set up from the same source population). However, artificially pairing samples where no biological rationale exists is not ideal.

The CMH-test as applied to genome-wide marker data is usually implemented in the popular package PoPoolation2 (Kofler et al., 2011), which aims to identify differences in allele frequencies that are consistent across biological replicates (Kofler

et al., 2011). However, this package does not account for heterogeneity between replicates and thereby confuses this heterogeneity for a main effect. For example, Table 2.1 shows a hypothetical contingency table with an inconsistent allele frequency difference by any reasonable definition, for which the CMH-test reports a significant result. This is surprising, because much of the rationale for using replicate lines in artificial evolution experiments is to distinguish genes that can be confidently identified as diverging due to selection rather than drift, as only the former should be consistent across lines. The genetic analysis tool PLINK (Purcell et al., 2007) also implements the CMH-test and while the documentation recommends testing for heterogeneity, this is not routinely done in published studies. At the time of writing, the PoPoolation2 package had been cited 21 times with respect to the CMH-test and PLINK's implementation of the CMH-test 170 times in "Google Scholar". While the PoPoolation2 package is never cited along with a test for heterogeneity, several of the studies citing PLINK also report tests for heterogeneity (e.g. Mero et al., 2010)

Table 2.1. A hypothetical set of contingency tables. The "A" allele frequency difference between treatment lines "TL1" and "TL2" are not consistent across the three populations. A CMH-test gives the following significant results: Chi-squared = 55.66, df = 1, $p < 0.0001$, Common Odds Ratio = 6.98.

		Allele	
Replicate	***Treatment Line***	***A***	***a***
1	1	66	5
	2	90	3
2	1	72	3
	2	60	5
3	1	69	21
	2	6	72

2.1.2 Examples of the CMH-test in the Literature

Recently the CMH-test has become highly popular in evolve and resequence (E&R) studies. Several such studies have considered data from a base population and 3

replicate treatment lines of *Drosophila melanogaster* sampled at various generations of experimental evolution under altered temperature regimes (Orozco-terWengel et al., 2012; Franssen et al., 2014; Kapun et al., 2014). Each generation, 500 females were sequenced by pool-seq, and the change in frequency between the base population and a given generation of each replicate determined. A CMH-test was used to test if the differences in allele frequencies between treatments were consistent across replicates (Orozco-terWengel et al., 2012; Franssen et al., 2014). These studies identified regions indicative of haplotype blocks under selection by finding consistent, large average changes in allele frequencies across replicate treatment lines in response altered temperature regimes (Franssen et al., 2014). Another study based on the same experimental evolution dataset used three replicates of two selection regimes (hot and cold) (Kapun et al., 2014). The authors used SNP frequencies within inversions to infer changes in inversion frequencies between the selection regimes. Consistency of inversion frequency differences across replicates was tested with the CMH-test. Allele frequencies were calculated for SNPs at multiple time points (Kapun et al., 2014). The study found significant, consistent changes in inversion frequencies between treatments across replicates, and the authors quantified the variation in changes of inversion frequencies. These studies do not report any attempts to test whether odds ratios across replicates are equal nor do they report how much allele frequency differences vary between replicates and they do not account for frequency variation that arises from coverage greatly exceeding the number of independent chromosomes in each pool, which is essentially pseudoreplication (Kolacskowski et al., 2011).

Another E&R study considered adaptation to viral infection rates. Four replicates of three regimes were compared; ancestral, sham-control and infected, where adult flies from each generation were either not pricked, pricked with a sterile needle or pricked with *Drosophila* C Virus (DCV) respectively (Martins et al., 2014). Allele frequencies were compared using a CMH-test and also compared to results using a Binomial GLM. This study does not report levels of variation in allele changes between replicates but found that results were the same for the two statistical tests used (Martins et al., 2014). Other examples of E&R studies that use the CMH-test to infer consistent allele frequency differences across replicates include response to novel laboratory environments (e.g. Huang et al., 2014).

In all the cases described above, changes in allele frequency are taking place from a common base population in response to particular or directed selection on a restricted number of traits that may have a relatively narrow range of mutational targets. Studies also use these statistical methods to find SNPs associated with naturally divergent traits such as coat colour in domestic horse breeds (McCue et al., 2012), pigmentation variation in wild populations of *D. melanogaster* (Bastide et al., 2013), as well as loci influencing economically important traits in GWAS-style analyses (Ayllon et al., 2015). The same approach can also be used in case-control studies to find disease risk loci, which is conceptually identical to finding consistent allele frequency differences between two or more groups (e.g. Mero et al., 2010; Cichon et al., 2011).

While the above studies have yielded many promising results there is nevertheless an issue with the application of the CMH-test which may result in numerous false positives. There is very seldom any attempt reported at assessing whether candidate SNPs found conform to the assumptions of the CMH-test, in particular the homogeneity of odds ratios. Such violations are likely to be common in many data sets and will produce false positives, which may be more frequent than true hits even after applying corrections for multiple testing. In fact, in a recent simulation study the CMH-test was found to have very low precision in identifying SNPs under selection (Topa et al., 2015). Guides to E&R studies maintain that the CMH-test performs better than many methods based on other simulations (Kofler & Schlötterer, 2014), though these, and other, simulations do not seem to consider the special cases of pool-seq designs (Baldwin-Brown et al., 2014; Kofler & Schlötterer 2014). Meanwhile other simulation studies do not consider a range of statistical approaches (Baldwin-Brown et al., 2014; Kessner & Novembre 2015) and a consensus over best practices is lacking (Kessner & Novembre 2015). Additionally, usually no attempt is made in such studies to correct for the violations of independent counts although such corrections have been suggested in other contexts (e.g. Kolacskowski et al., 2011; Bergland et al., 2014; Machado et al., 2016).

2.1.3 Binomial Generalised Linear Models (GLMs), Quasibinomial GLMs (QBGLMs) and Linear Models (LMs)

Another approach is to model allele frequencies in a GLM with binomial error

distributions (binomial GLMs). This approach estimates the effects of a trait of interest, population of origin as well as their interaction on the allele or read count. This is similar to approaches that identify differential expression in RNA-sequencing (RNA-seq) experiments (Lund et al., 2012; McCarthy et al., 2012). Examples of binomial GLMs are less common in the literature to infer consistent allele associations with a stratum across population, although Martins et al., (2014) report using binomial GLMs to compare results with the CMH-test. Binomial GLMs have been used to analyse allele frequencies in other contexts (e.g. Bergland et al., 2014; Jha et al., 2015; Machado et al., 2016; Kapun et al., 2016). A related statistical framework is the GLM with quasibinomial error distribution (quasibinomial GLM) that includes an extra parameter, φ, that can account for variation over and above that assumed by a binomial distribution (Crawley 2013). Finally, a linear model is also possible where the allele frequencies are modeled as frequencies with the treatment group as a dependent variable. The latter two approaches have the added benefit that they need not assume a specific pairing of an experimental treatment line with a "control" or other line but a pairing can be added if there are good biological reasons to do so.

2.1.4 The G-test

The G-test is not commonly used in population genomics and is also based on the analysis of multi-way contingency tables. The G-test is related to the Chi-squared test but with more general application. The G-test is based on the log-likelihood ratio test, which is approximated at large sample sizes by the common Chi-squared test (Sokal and Rohlf 1969). The G-test is less reliable when cell counts in the tables are 0 (Sokal & Rohlf, 1969, 1981) though continuity corrections where cell frequencies are low can make the test more robust (Sokal & Rohlf, 1969, 1981). To my knowledge the G-test has not been applied to the analysis of experimental evolution studies though it has been used in the analysis of genomic data in QTL studies using bulk segregant analysis (Magwene et al., 2011).

In summary, the CMH-test, which has become a popular method of testing for consistent allele frequency differences across biological replicates of stratified populations, is only valid under specific assumptions if it is to be used to detect allele

frequency differences in stratified populations. Odds ratios across replicates should be homogenous (e.g. no drift, no founder effects or idiosyncratic responses to selection), and data contributing to cell counts should be independent. These assumptions are frequently violated in real datasets and apparently this is not commonly tested for. While violations of these assumptions are not fatal to the test in general, inferences about the consistency of allele frequency differences based on the results of this test are only reliable if these assumptions are satisfied. The aim of this study is to highlight the problems arising from a failure to test whether these assumptions hold and to assess, by simulation, the performance of different methods to identify consistent differences in allele frequencies between two treatment groups across biological replicates.

2.2 Methods

All simulations and analyses were performed in the R programming language (R Development Core Team, 2014). All code to run and analyse the simulations, including custom written functions, is available as a repository of code at the digital repository GitHub from https://github.com/RAWWiberg/ER_PoolSeq_Simulations. A script for applying a QBGLM to a real dataset, as in the re-analysis of the Orozco-terWengel *et al.,* (2012) data, is available from https://github.com/RAWWiberg/poolFreqDiff.

2.2.1 Description of Simulation Protocol and Parameter Value Choice

The behaviour of the G-test, CMH-test and Binomial GLMs are explored using simulated datasets. 1,000,000 (of which 1% [10,000] were designated "true positives", see below) independent number sets that represent biallelic SNPs are generated across k replicates of two treatment lines assuming a simple but realistic population genetic model that reflects a standard experimental evolution design. The neutral or "null" case of an experimental evolution scenario can be described as an instantaneous fission model where k replicate subpopulations originate from a common ancestral population. Replicates are then split into two "treatment lines" which evolve by drift for t generations leading to some differentiation ($F_{ST} > 0$) from the ancestral population. I assume that the ancestral population is not under selection and is at drift-mutation equilibrium. Thus, the "A" allele frequency in the ancestral population (p_A) is drawn

from a beta distribution B(α, β) where:

$$\alpha = 4Neu$$

and,

$$\beta = 4Nev,$$

where u is the forward mutation rate and v is the backward mutation rate between the two states of a biallelic SNP respectively and Ne is the effective population size (Charlesworth and Charlesworth 2008). The "A" allele frequency within each "treatment line" (f_A) is generated as a sample from a truncated normal distribution bounded between 0 and 1 (Nicholson et al., 2002; Balding 2003) with mean $\mu = p_A$ and variance $\sigma^2 = F_{ST}p_A(1-p_A)$, where F_{ST} represents the amount of neutral divergence from the ancestral population (F_{ST}) after t generations (Nicholson et al., 2002; Balding 2003). No SNPs are allowed to become fixed for the same allele across all lines.

Finally, because the entire population is rarely analysed in experiments a sample allele frequency at each locus of size $2N = n$ alleles is drawn from each treatment line using the binomial distribution B(n, f_A) so as to obtain the count (x) of the "A" alleles in the sequenced pool. The count of "a" alleles in the pool is then n-x and the frequency of "A" in the pool (f_{Apool}) is x/n.

Data are generated by progressively filling the cells of contingency tables. Each partial table represents a separate replicated pair of experimental evolution lines. The rows within a table give the frequencies of the alleles at a single SNP for one "treatment line". Total row counts (CT) are either sampled or fixed. The allele counts (C_A and C_a) within the lines are then calculated by:

$$C_A = f_{Apool} * CT$$

and,

$$C_a = CT - C_A$$

In pool-seq data allele counts are commonly derived from raw read counts at each locus. This can lead to substantial variation in CT across genomic regions, and treatment lines due to differences in sequencing efficiency or random variation in

coverage. That average coverage is often greater than the number of chromosomes in the pool should be called pseudoreplication (Kolacskowski et al., 2011; Feder et al., 2012). The double sampling nature of pool-seq has been recognised and ways to deal with it have been proposed (e.g Lynch et al., 2014). One way to ameliorate effects of this is to rescale counts to correspond to frequencies out of the known number of chromosomes in the sample or to a computed effective sample size (n_{eff}) (Kolacskowski et al., 2011; Feder et al., 2012; Bergland et al., 2014). Here, simulation results are compared where *CT* varies uniformly between 16 and 400 or is fixed at 100, 200 or n_{eff}. For the purposes of the n_{eff} correction, read depth is *CT* and sampled as above, the number of chromosomes/alleles in the pool (*n)* is 2*N* and n_{eff} is given by:

$$n_{eff} = ((n*CT)-1)/(n+CT)$$

according to Kolacskowski et al., (2011) and Feder et al., (2012).

To parameterise the distributions in these simulations, it is necessary to take realistic values for the various population parameters. For mutation rate (*u* and *v*) values on the order of between $2x10^{-9}$ and $1x10^{-8}$ are common in e.g. *Heliconius melpomene* (Keightley et al., 2015) or *D. melanogaster* (Haag-Liautard et al., 2007; Keightley et al., 2014) and estimates of N_e reported are on the order of 1,000,000 - 2,000,000 in these and other species (Charlesworth and Charlesworth 2008; Keightley et al., 2014; Keightley et al., 2015). Thus, the parameters of the Beta distribution describing the allele frequencies in the base population are $4N_eu = 4N_ev = 0.2$. Several experimental evolution studies have recently been published (see Introduction). Many of these studies represent evolution over relatively few generations and few of them report standard population genetic divergence statistics. Nevertheless, when these data are available, fairly substantial F_{ST} estimates are typically reported. Kang et al., (2016) report estimates of F_{ST} between 0.08 and 0.2 after ~50 generations of experimental evolution. Meanwhile, even after only 3 generations of evolution by drift Santos et al. (2013) report differentiation of between 0.002 and 0.012. Some experimental evolution studies have been run for many more generations (~100 generations: Immonen et al., 2014), in which case even higher estimates of F_{ST} are expected (~0.3-0.5). Neutral differentiation (F_{ST}) will also depend on the population size. Here I simulate data using values of 0.1,

0.2, and 0.3 for F_{ST} , which is probably conservative. I assume a pool size (N) of 100 throughout which is on the same order of magnitude as other experimental evolution studies (e.g. Orozco-terWengel et al., 2012; Martins et al., 2014) and of recommended sample sizes (Schlötter et al., 2014).

The primary aim of this study is to assess the False Positive Rates (FPRs) of different statistical tests. Under a null hypothesis a well-behaved statistical test should produce a uniform distribution of p-values ranging from 0 to 1 (Storey 2002; Story and Tibshirani 2003). Thus, for a given cut-off threshold α, the proportion of tests with a p-value <= α should be α. This can be represented as a straight 1-1 line of the FPRs at different values of α against α on a log-log plot. To evaluate the statistical tests in this study, the FPRs for α = 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, and 0.5 is calculated for each test. The simulations are run to consider k = 2, 3, 4, and 10 replicates. The CMH-test, CMH-test+Woolf-test, binomial GLMs, quasibinomial GLMs, the G-test, as described in Sokal and Rohlf (1969; 1981), and a Linear Model (LM) are then applied to each simulated SNP. Because the allele frequencies produced in these simulations are random draws and the population genetic model applied is a neutral one, the simulations represent a null or "neutral" scenario and most simulated SNPs are expected to show no consistent difference across the k samples.

While the main aim of this study is to evaluate the FPRs of these statistical tests, the power of the tests is also of interest. Thus, 1% of the 1 million SNPs (10,000 SNPs) are designated "true positives." For each true positive the SNP frequencies are simulated as above with the exception that one treatment is given a consistent allele frequency increase of 0.2 on top of whatever change is generated by drift. Power can then be roughly assessed by estimating the proportion of all true positives recovered among bottom 1% of SNPs of the p-value distribution.

2.2.2 Implementation of the CMH-test

CMH-tests are performed using the R function mantelhaen.test() from the "stats" package. This same function is used in the popular software package PoPoolation2 (Kofler et al., 2011). Heterogeneity was tested using a Woolf-test (Woolf, 1955) from the same R package. Counts of zero are tolerated by the CMH-test but not by the Woolf-test where a common procedure is to add one to each zero count cell. Here one is

added to all cells if any cells have a count of zero for both the Woolf-test and the CMH-test. In the CMH-test framework, a consistent result should be one that shows a common odds ratio significantly greater than one as well as a non-significant test of heterogeneity in odds ratios.

2.2.3 Implementation of Binomial GLMs, Quasibinomial GLMs and LMs

GLMs are run in the standard R glm() function, from the "stats" package. Two model structures are tested for binomial GLMs:

$$(1): y = treatment + replicate + treatment{:}replicate + e$$

and,

$$(2): y = treatment + e$$

Where y gives the counts of "A" and "a" alleles, treatment, replicate, and treatment:replicate are the treatment, replicate, and interaction effects respectively. e is a binomially distributed error term. A consistently associated SNP is one where there is both no evidence for a 2-way interaction between treatment line and replicate on allele frequency (LxR interaction) and an overall significant effect of treatment line (L) on allele frequency, this is tested by model structure (1). Model structure (2) simply tests whether there is an overall effect of treatment. Inconsistent allele frequency differences should increase variance in one treatment and give non-significant treatment effects. Under model structure (1) p-values for the treatment and interaction effects are obtained from likelihood ratio tests. For model structure (2) p-values are from t-tests. Counts of zero are tolerated by the GLM but can lead to other problems due to fitted values from the link function being undefined. To counter this, a common procedure is to add a count of one to each allele count if any zero counts are encountered, which I adopt here. Quasibinomial GLMs are also fitted with the glm() function (family="quasibinomial"). Only the model structure (2), see above, is tested because there are not enough residual degrees of freedom to test for interaction effects. Interaction effects are estimated for binomial GLMs because dispersion is assumed to be 1. However, these estimates should be treated with a degree of caution. For quasibinomial GLMs, e is a quasibinomially distributed error term and p-values for the treatment effects are obtained from t-tests.

Finally, a general Linear Model (LM) is implemented with model structure (2) in the function lm(). In the LM, e is the Gaussian error term. P-values for the treatment effects are obtained by t-tests.

2.2.4 Implementation of the G-test

G-tests are performed as described in (Sokal & Rohlf, 1969) using a custom written R function. In a G-test framework a SNP allele that occurs at consistently different frequencies between lines across populations is one which shows an overall association between allele and line (LxA) as well as a non-significant line by population by allele count interaction (LxAxP interaction). Counts of zero are not tolerated by the G-test. Again, a common procedure is to add 1 to all cells if any cells have a count of zero, this procedure is applied here.

All simulations and analyses were performed in the R programming language (R Development Core Team, 2014). All code, including custom written functions are available at: https://github.com/RAWWiberg/ER_PoolSeq_Simulations. Data presented below are archived in the Dryad repository: http://dx.doi.org/10.5061/dryad.mn0tv

2.2.5 Re-Analysis of a Dataset

Data from the E&R study on adaptation to novel temperature environments in *D. melanogaster* is re-analysed (Orozco-terWengel et al., 2012). Raw data files, as generated by the PoPoolation2 package, are available from the Dryad online repository (Orozco-terWengel et al., 2012; http://dx.doi.org/10.5061/dryad.60k68.2). These data are re-analysed using the unpaired quasibinomial GLMs. A pipeline for preparing this data for the re-analysis described below is available a repository of code on GitHub (https://github.com/RAWWiberg/ER_PoolSeq_Simulations). The original data analysis is described in Orozco-terWengel et al., (2012), and also re-analysed in Topa et al., (2015) and in Iranmehr et al., (2016). Here I compare the results from the original study and re-analyse the raw data with some modifications. The full dataset contains 1,547,837 SNPs from six pools of 500 individuals each. I consider only truly biallelic SNPs, as in Topa et al., (2015). The minimum and maximum coverage thresholds remain as in Orozco-terWengel *et al.,* (2012) (min-count = 10, min-cov = 10, max-cov = 500). Analyses are performed on the raw allele counts and counts that are re-scaled to

be out of either 1,000 (to match the number of independent chromosomes in the pool), 100 or n_{eff}. In total, 1,370,371 SNPs are analysed. The base (B) and 37th generation (F37) from the experiment are compared across three replicated experimental evolution lines in order to identify consistent allele frequency differences between the two generations across the three replicates.

2.3 Results

2.3.1 Simulated Dataset

The distributions of the mean allele frequency difference between the lines and the standard deviation (SD) of these differences are shown in Figures 2.1-2.3. The SD can be viewed as a measure of how consistent the difference between the two treatment groups is. The SD is inexact since its calculation requires a pairing of treatment lines while some statistical tests do not assume a pairing and in many experimental designs no meaningful pairing exists. There is no systematic relationship between the mean allele frequency difference and the SD of allele frequency differences indicating that these are varying quite freely in the simulations. Because the results are qualitatively the same for all values of F_{ST} only data from simulations using a value of F_{ST} = 0.2 are presented below.

2.3.2 False Positive Rates

There is substantial variation in the False Positive Rates of each of these tests (Figure 2.4). It is clear that the FPRs for the CMH-test are seriously overinflated even at very stringent values of α. This pattern is particularly notable where allele frequencies are given by the raw allele counts which are allowed to vary (Figure 2.4). The FPRs are also highly inflated even when the Woolf-test is used in an attempt to identify SNPs where the odds ratios are not homogenous across the partial tables (Figure 2.4). Similarly, the FPRs are high for the G-test, as well as Binomial GLMs. In contrast, GLMs with a quasibinomial error distribution and the regular LM show FPRs that are more appropriate. Both approaches (the LM and quasibinomial GLMs) produce FPR lines that lie very close to the expected 1-1 line (Figure 2.4). The largest inflations of FPRs are again seen in simulations where the allele counts are allowed to vary and are

very low in the simulations where allele counts are fixed at 100, 200 or scaled to the effective sample size (n_{eff}) (Figure 2.4). In terms of the FPRs it is clear that the quasibinomial GLMs and the LMs perform best.

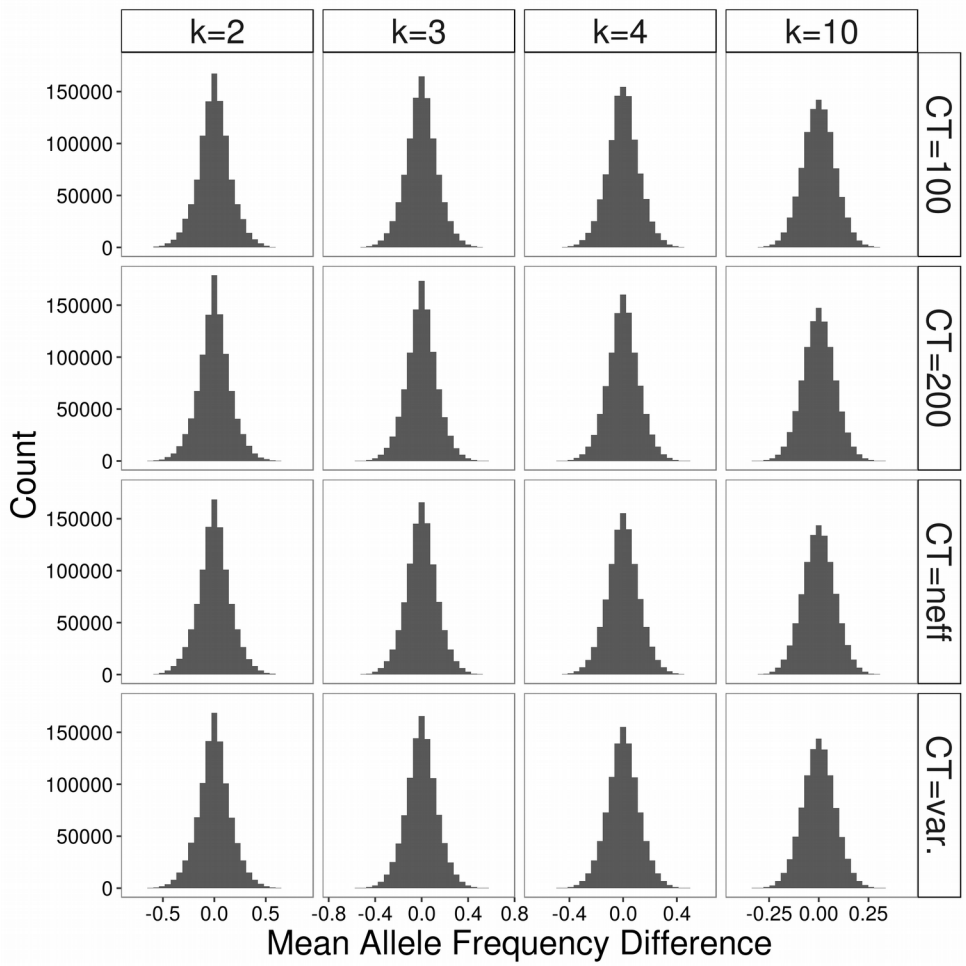


Figure 2.1. Distribution of mean allele frequency differences between treatment lines across replicates for each of the simulations. Data are shown for simulations that consider k = 2, 3 and 4 replicates. "CT var." - row totals in the 2-way tables can take any value between 16 and 400, "CT = 100" - row totals in each of the 2-way tables are fixed at 100, "CT = 200" - row totals in each of the 2-way tables are fixed at 200, "CT = neff" - row totals in the partial tables are scaled to the effective sample size. Only the "neutral" SNPs are shown.

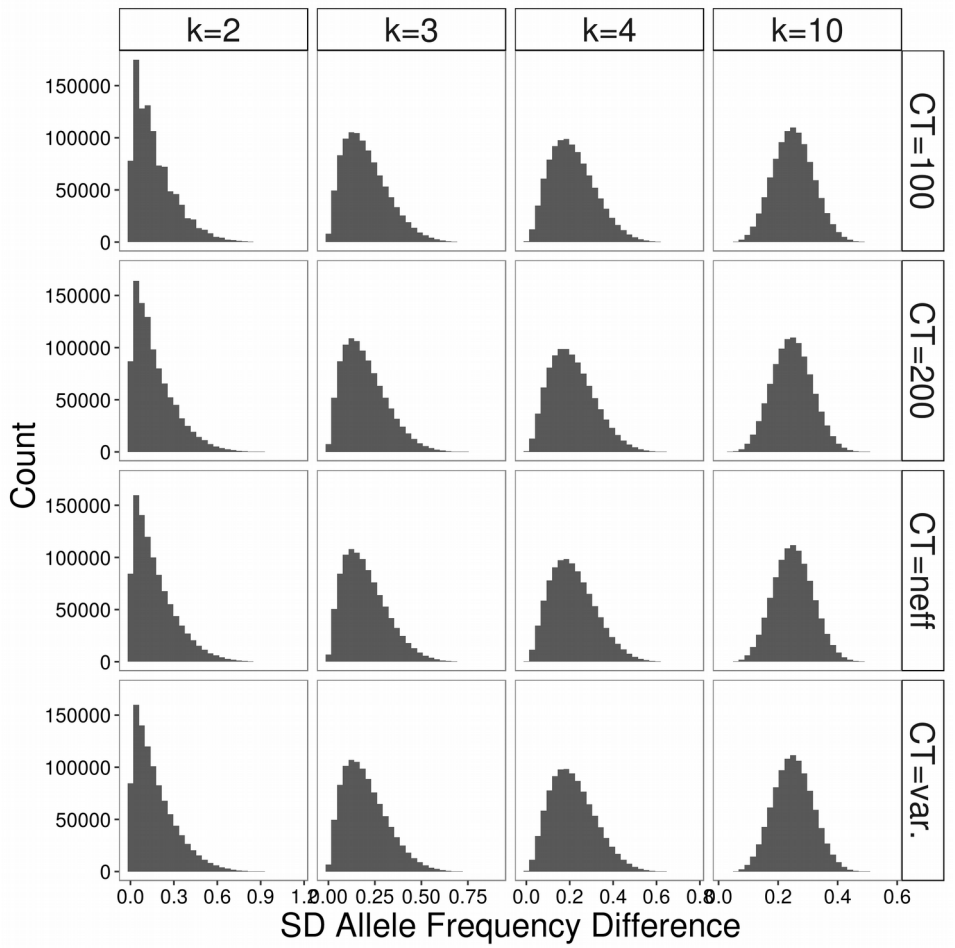


Figure 2.2. Distribution of standard deviation (SD) of allele frequency differences between treatment lines across replicates for each of the simulations. Labels are as in Figure 2.1.

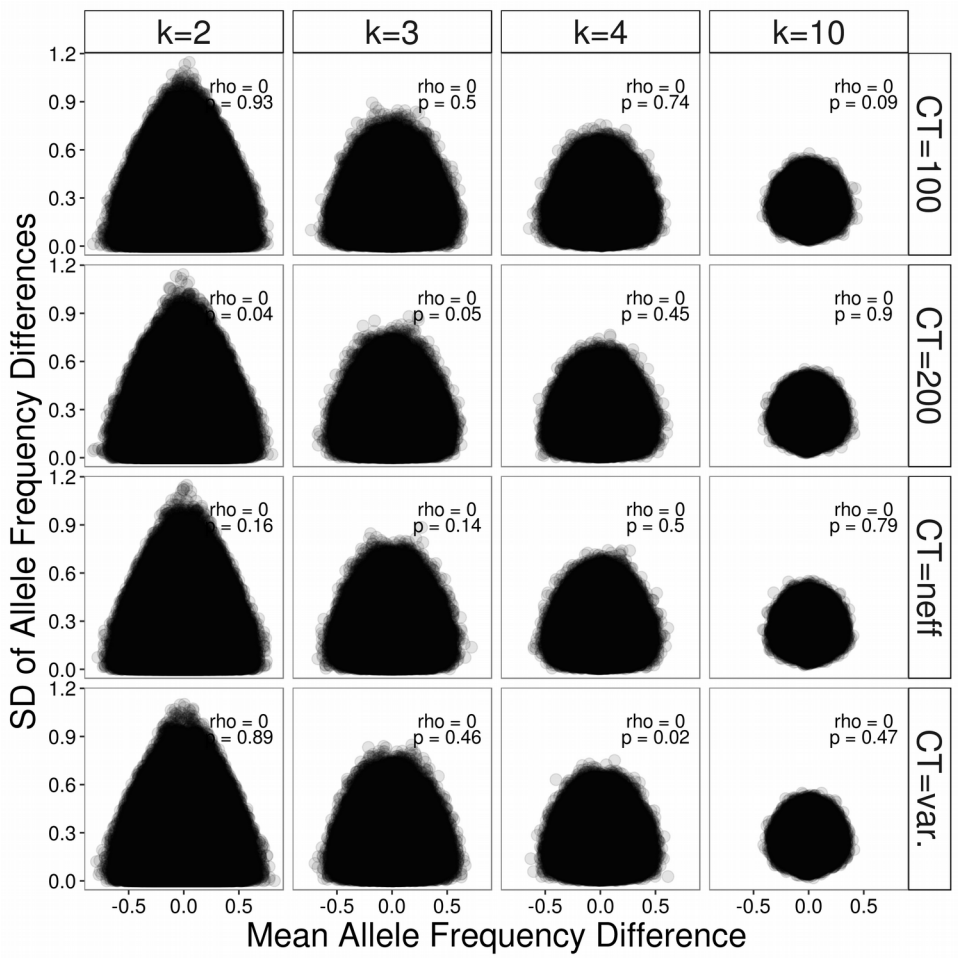


Figure 2.3. Relationship between the mean difference in allele frequencies between treatment lines, across replicates and the standard deviation (SD) of allele frequency differences. Labels are as in Figure 2.1. Inset text gives the p-values and correlation coefficients (rho) for Spearman Rank correlation tests between the x and y variables.

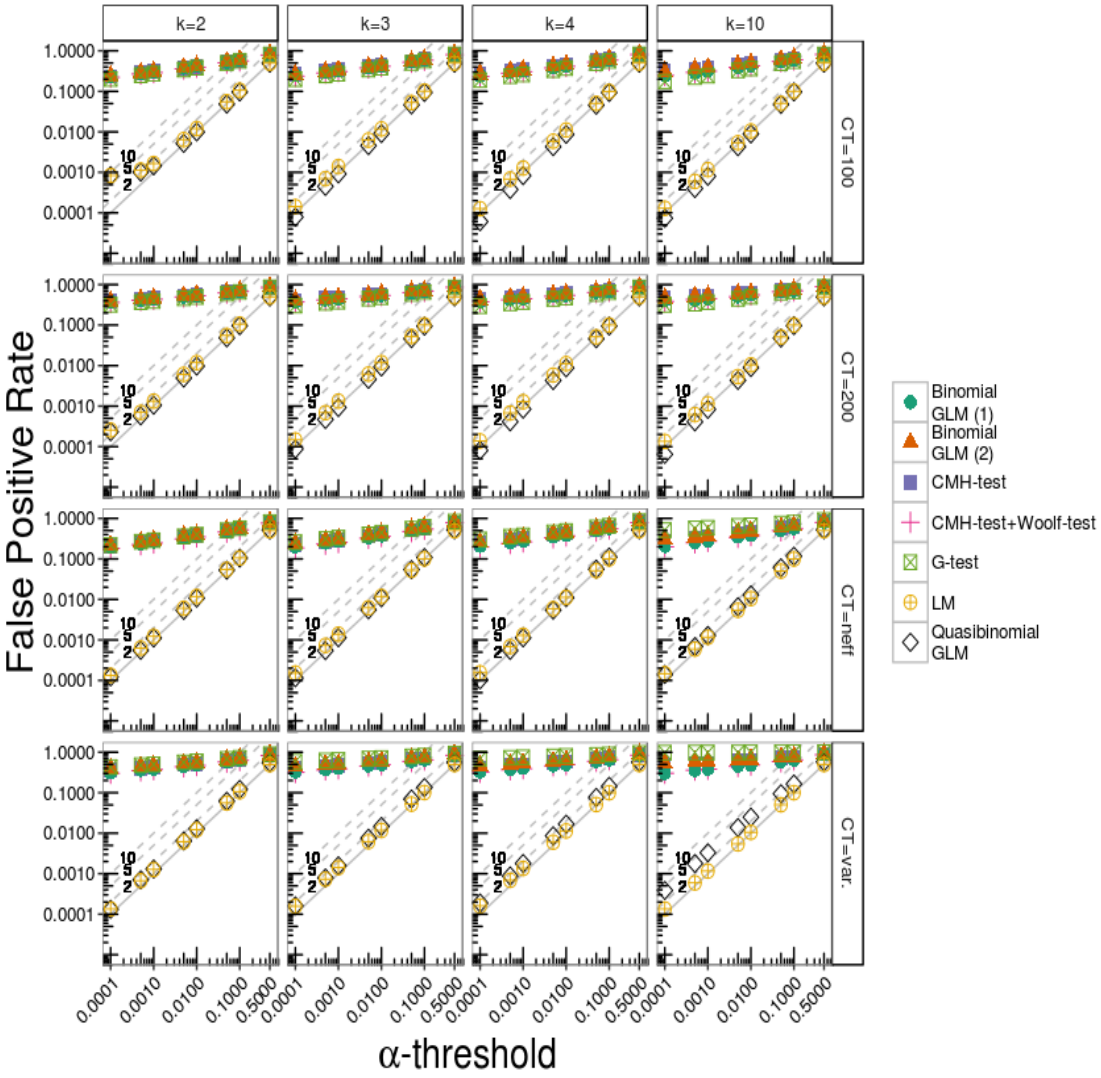


Figure 2.4. The False Positive Rate (FPR) at different levels of a for each simulation. Simulations are run for k = 2, 3, 4 or 10 replicatd treatment lines to a neutral FST of 0.2. Allele counts are allowed to vary freely (CT=var.), are fixed at 100 or 200 (CT=100, CT=200) or are scaled to the "effective sample size" (CT=Neff). The diagonal lines labelled "2", "5" and "10" represent 2, 5, and 10-fold inflations of p-values respectively.

2.3.3 True Positive Rates

These simulations also implemented a simple method for assessing the power of the different tests. The True Positive Rate (TPR) is calculated as the proportion of all true positives that were seeded in the simulations that are recovered among the SNPs below the 1st percentile of the p-value distributions (hereafter the "top 1% of SNPs"). In general the CMH-test seems to perform quite well recovering between ~20 and 29% of true positives. However, quasibinomial GLMs and LMs perform better, recovering ~30% of true positives among the top 1% of SNPs (Figure 2.5). The remaining statistical tests (Binomial GLMs and G-tests) perform rather poorly recovering less than 5% of true positives. While there are some differences in the TPRs as the allele counts are allowed to vary or kept fixed, the TPR is primarily influenced by the number of replicates (Figure 2.5). It should be noted that the precise TPRs will vary with how large the average difference between treatment lines due to selection is in comparison to neutral differentiation among the treatment lines. In this simulation the value added consistently to one treatment as a difference due to selection was 0.2. Indeed, when simulations are run using $F_{ST} = 0.1$ the TPR is higher in all cases. Thus, these values should be taken as a guide and other scenarios, including variation in the average difference between treatment lines due to selection among SNPs, might produce different results. However, the distribution of TPRs from multiple simulations with the same parameters is narrow, especially for simulations of 1,000,000 SNPs (Figure 2.6).

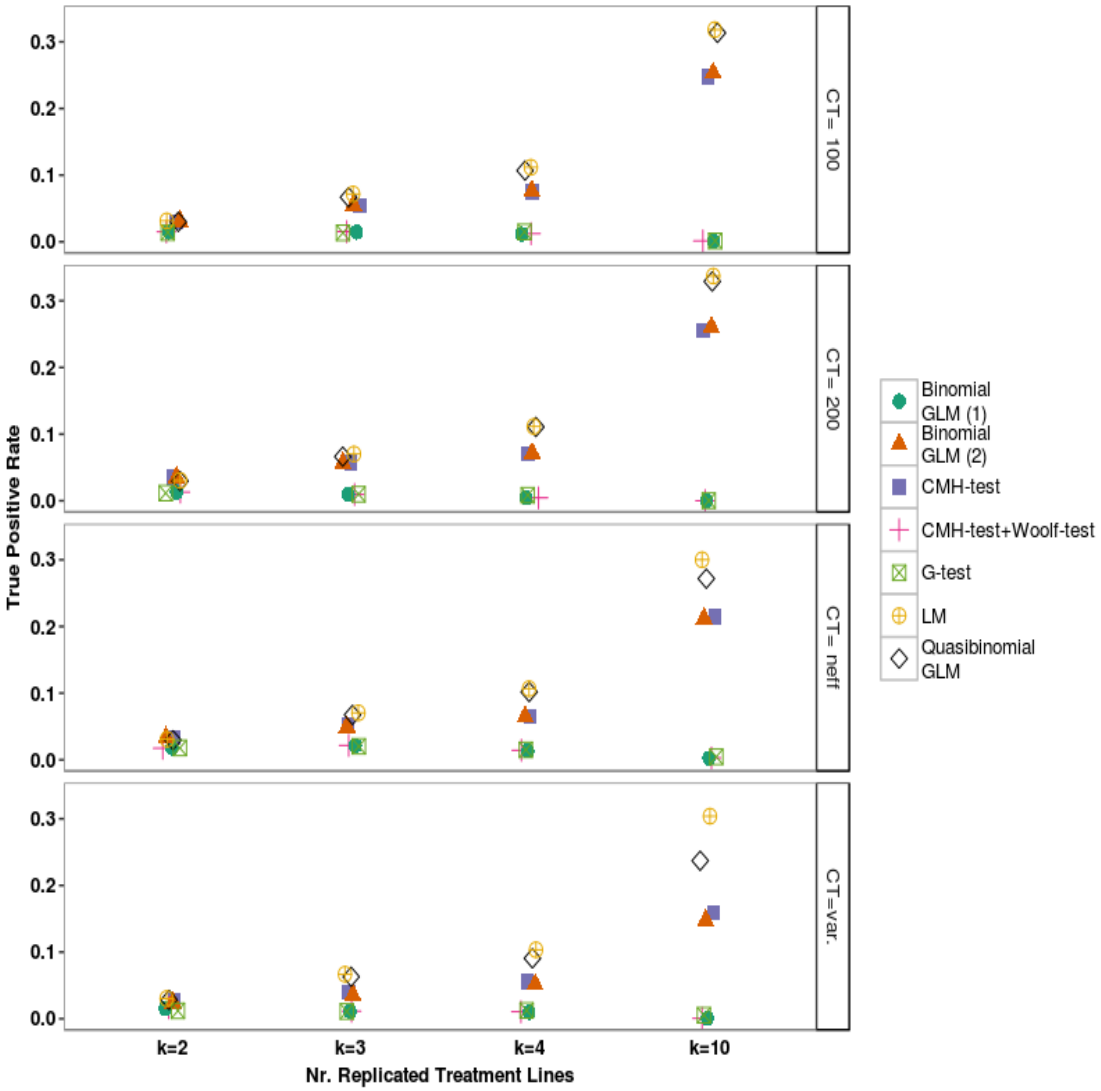


Figure 2.5. The True Positive Rate (TPR). TPR is calculated as the proportion of all true positives that are recovered among the top 1% of SNPs. Simulations are run for k = 2, 3, 4, and 10 replicated treatment lines to a neutral FST of 0.2. The selection differential applied to true positives is 0.2. There are 10,000 true positives in each simulation. Allele counts are allowed to vary freely (CT=var.), are fixed at 100 or 200 (CT=100, CT=200) or are scaled to the "effective sample size" (CT=neff). Points are "jittered" horizontally to avoid overlap.

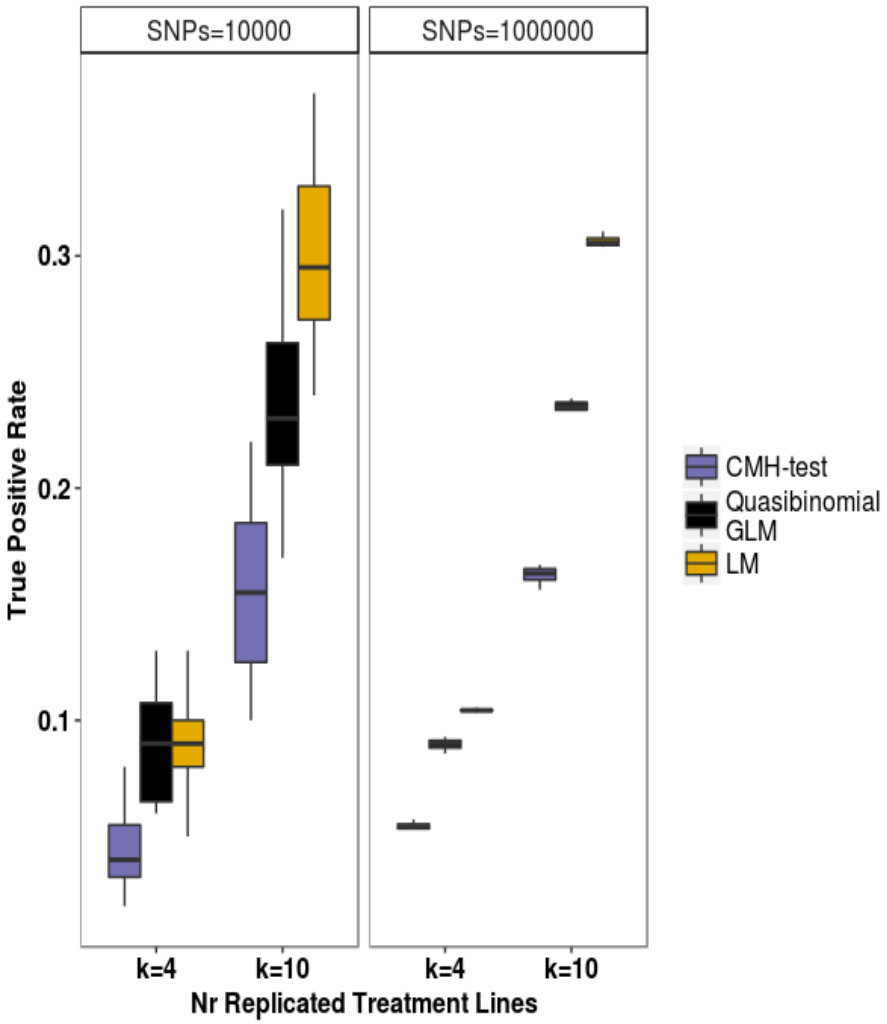


Figure 2.6. The consistency and scaling of the True Positive Rate (TPR) across simulations. Results are shown for 10 repeated simulations of 10,000 SNPs at k = 4 or 10 (left panel), or 4 repeated simulations of 1,000,000 SNPs at k = 4 or 10 (right panel). In all simulations allele counts are scaled to be out of 100, neutral divergence (F_{ST}) is set to 0.2, and the a difference between treamtnet lines applied to simulate an average difference due to selection is 0.2, as in the main results (Figure 2.5). Only results for the CMH-test, quasibinomial GLMs and LMs are shown.

2.3.4 Re-analysis of Dataset

The analysis of allele frequencies from raw counts produces somewhat similar results to the original analysis (Orozco-terWengel et al., 2012) (Figure 2.7). Spurious false positives due to excessive coverage near chorion gene clusters on chromosome 3L (Orozco-terWengel et al., 2012) are no longer apparent (Figure 2.7). However, scaling counts to match the large number of chromosomes in the pools (to be counts out of either 100 or 1,000) produces unusual looking Manhattan plots (Figure 2.8), likely because it creates artificially high confidence in the measurements within the quasibinomial GLM resulting in inflated -log10(p-values). A random sample of 100 of the SNPs that are significant after Bonferroni correction suggest that these high scoring SNPs still show patterns that researchers would want to identify, i.e. they show a consistent difference between the two time points across replicates (Figure 2.9 and Figure 2.10). Using raw allele counts or scaling counts to correspond to n_{eff} does not produce this inflation (Figure 2.7).

Because quasibinomial GLMs produce the expected uniform distribution of p-values under the null hypothesis (Figure 2.4), it is possible to apply standard corrections for multiple testing. The number of SNPs that achieve genome-wide significance using q-values (Storey 2003; Storey and Tibshirani 2015), Benjamini-Hochberg (B-H) (Benjamini and Hochberg 1995), or Bonferroni correction are shown in Table 2.2. It is apparent that raw counts and counts scaled to n_{eff} are more conservative estimates at least for the Bonferroni correction. Methods that control the False Discovery Rate (FDR) (Q-values and B-H correction) are far more liberal and produce more "significant" SNPs (Table 2.2).

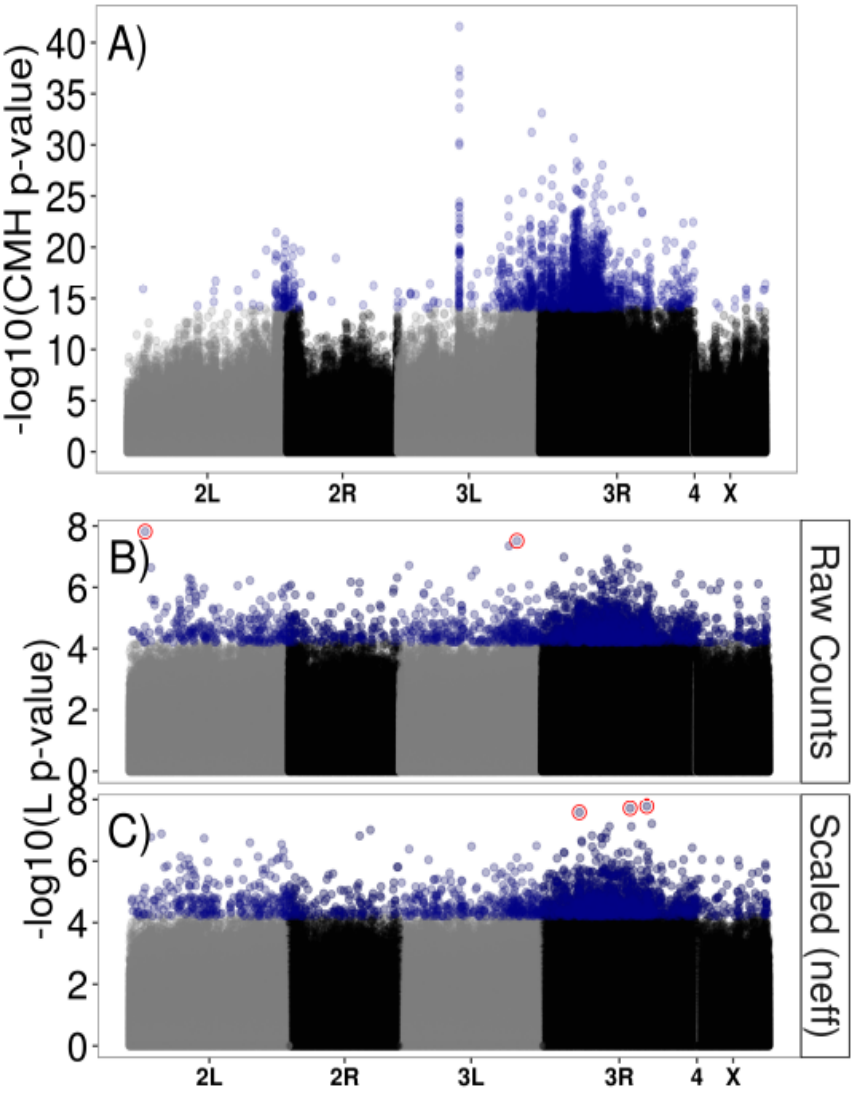


Figure 2.7. A) Manhattan plot of the original CMH-test results from Orozco-terWengel et al., (2012). Blue points are the top 2,000 SNPs identified by the CMH-test. B) and C) Manhattan plots of the re-analysis of the Orozco-terWengel et al., (2012) data using quasibinomial GLMs using the raw counts [B)] as well as scaling counts to n_{eff} [C)]. Shown are -log10(p-values) from the main treatment line (L) effect. Blue points are the top 2,000 SNPs. Red points are SNPs that pass genome-wide Bonferroni correction. Note the differences in the range on the y-axis between.

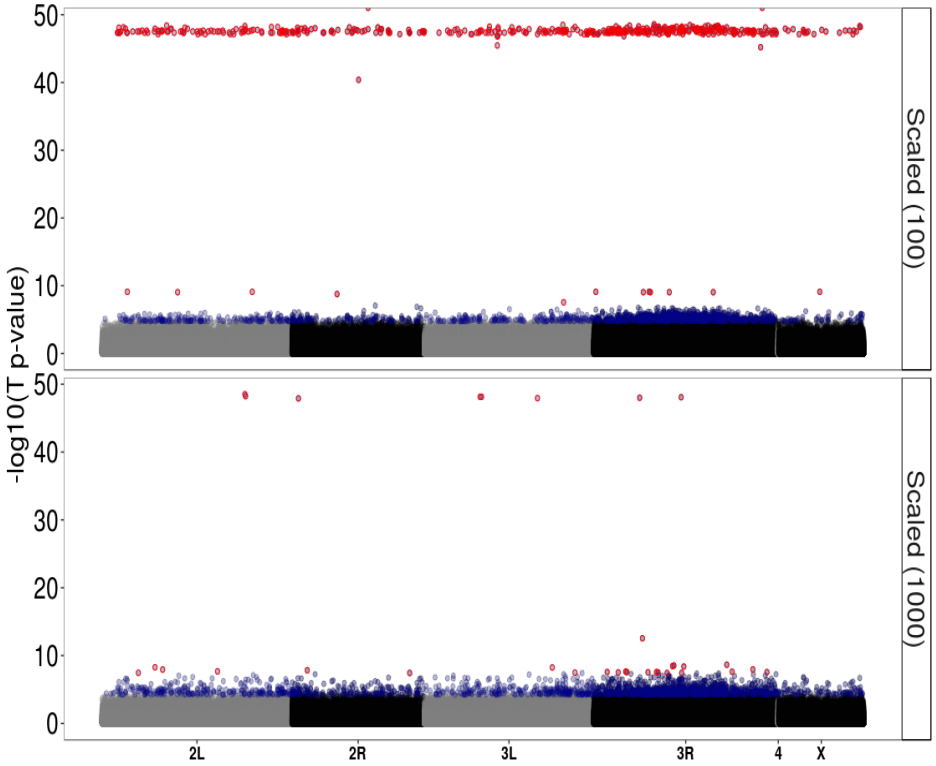


Figure 2.8. Manhattan plots of the re-analysis of the Orozco-terWengel et al., (2012) data using quasibinomial GLMs scaling counts to be out of 100 or 1,000. Shown are -log10(p-values) from the main treatment line (L) effect. Blue points are the top 2,000 SNPs. Red points are SNPs that pass genome-wide Bonferroni correction. Note the differences in scale on the y-axis.

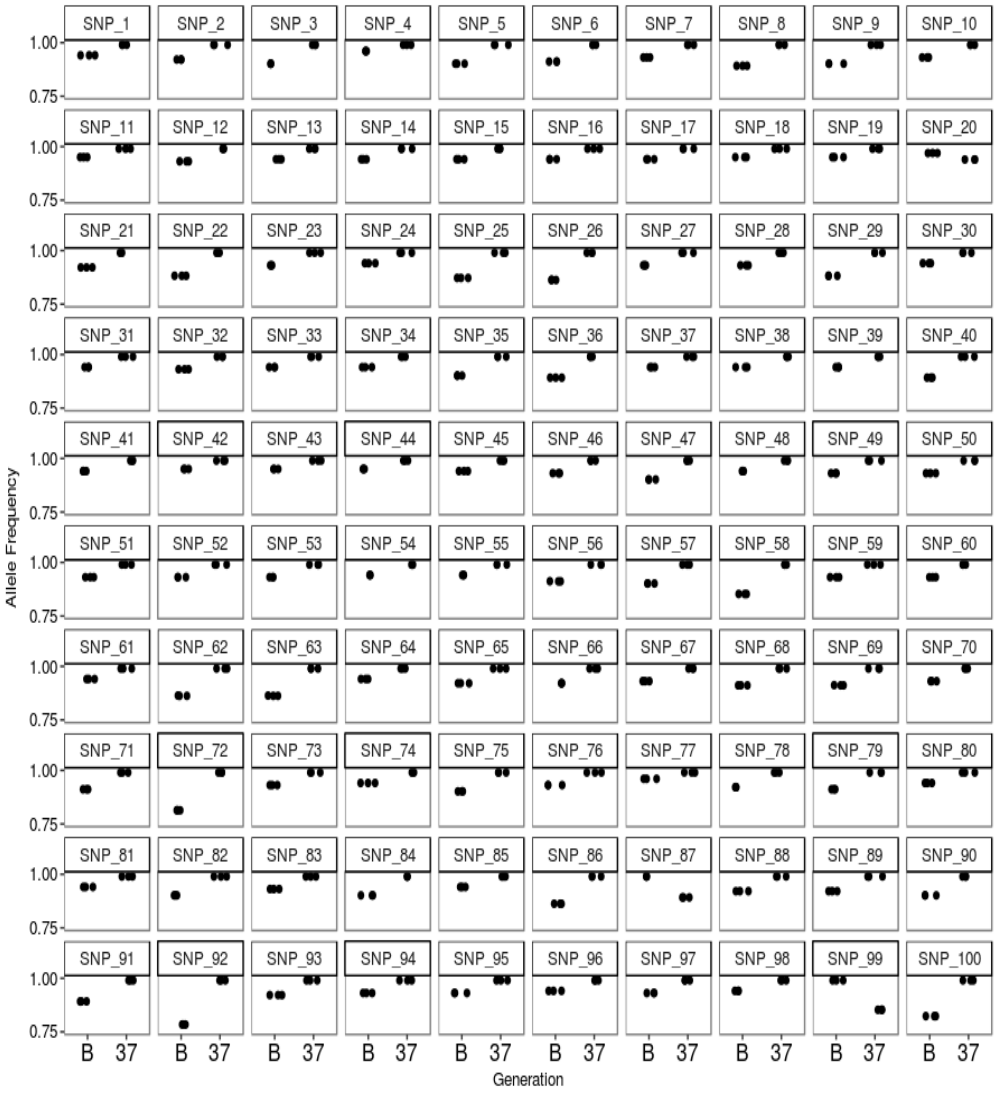


Figure 2.9. A random sample of 100 SNPs from the SNPs that pass genomewide Bonferroni signficance in the re-analysis of the Orozco-terWengel et al., (2012) dataset where allele counts have been scaled to be out of 100 (Figure 2.8). Points have been horizontally "jittered" to prevent overlap.

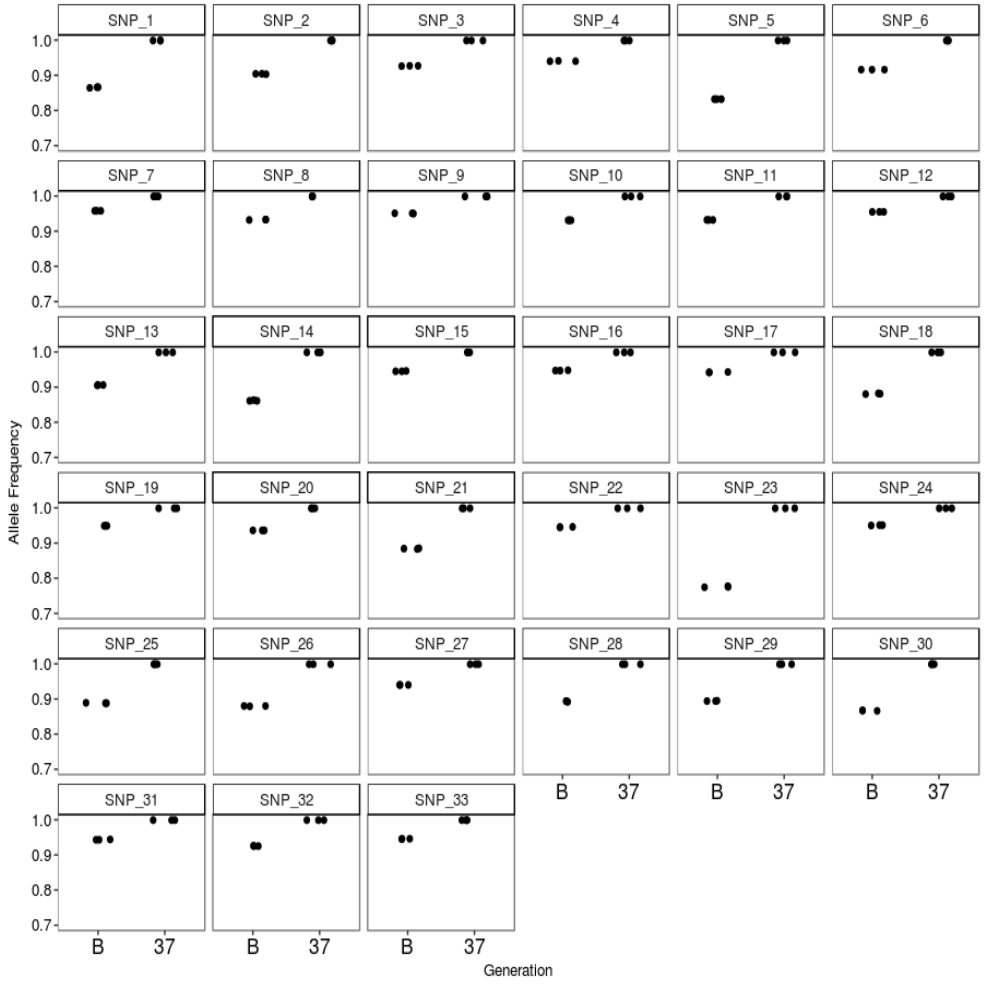


Figure 2.10. The 33 SNPs that pass genomewide Bonferroni signficance in the re-analysis of the Orozco-terWengel et al., (2012) dataset where allele counts have been scaled to be out of 1,000 (Figure 2.8). Points have been horizontally "jittered" to prevent overlap.

Table 2.2. The number of SNPs that pass multiple test correction in the re-analysed datasets. For Bonferroni correction the α threshhold 0.05 was divided by the total number of tests (SNPs tested) to get the genome-wide multiple test correction threshold. For Q-values and Benjamini-Hochberg (B-H) correction, a False Discovery Rate (FDR) threshhold of 0.05 was used. Bonferroni corrections carried out manually, B-H corrections followed the procedure in Benjamini and Hochberg (1995), and q-values were calculated using the "qvalues" package in R (Storey et al., 2015).

Re-analysis	***Bonferroni***	***Q-values***	***B-H***
Raw counts	2	67,702	3,961
Counts scaled to n_{eff}	3	67,505	4,571
Counts scaled to 100	456	61,022	15,053
Counts scaled to 1,000	33	13,013	2,532

2.4 Discussion

With the increasing popularity of pooled-sequencing methods to study population genomics and E&R studies, the importance of determining best practice statistical methods for allele frequency estimation and the identification of consistent allele frequency differences is crucial. User-friendly software packages remove the need for complicated scripting but also make statistical tests and analytical methods less transparent. This has led to several cases where statistical tests have been applied which do not address the question. The current study highlights problems with the way in which the popular CMH-test is applied and proposes some alternative methods.

The CMH-test produces a large number of significant test results under the null hypothesis and as such has very high False Positive Rates (FPRs) even at relatively high α thresholds. This seems to be because it very easily confuses heterogeneity for a main effect. Indeed, this potential is noted in much of the original literature describing this test (Landis *et* al., 1978; Agresti, 1996). Similarly, many other statistical tests assessed here have FPRs that are unacceptably high (G-tests and binomial Generalised Linear Models; GLMs). However, Linear Models (LMs) and quasibinomial GLMs perform very well under the null hypothesis producing uniform p-value distributions and the

characteristic 1-1 relationship, on the log-log scale, between the FPR and different thresholds of α.

Meanwhile the True Positive Rate (TPR), the ability to identify SNPs that are in fact under selection (true positives), varies across the tests and simulations. Quasibinomial GLMs and LMs perform the best, recovering more true positives than other statistical tests. In addition, there is a strong relationship between the number of replicated treatment lines in the experiment and the TPR regardless of the statistical test.

In addition to low FPRs and high TPRs, there are other attractive properties that the quasibinomial GLMs have over the CMH-test and the G-test. First, properly behaved p-values allow controlling False Discovery Rates (FDRs) e.g. by q-values, Bonferroni or Benjamini-Hochberg correction (Story and Tibshirani 2003). This is preferable to relying on arbitrary cut-offs of e.g. "the top 1%" or the "top 1,000" SNPs. Second, an attractive feature of quasibinomial GLMs over tests like the CMH or the G-test is that there is no need to arbitrarily pair experimental treatments. However, the option exists if it makes biological sense and if the data allow it.

The simulations in this study highlight an additional problem. When pools of individuals are sequenced, the coverage can vary substantially between pools or between genomic regions. This variation in coverage translates to differences in the total count for a SNP position. Our results indicate that variation in these counts between loci affects the performance of some statistical tests. A simple solution is to rescale all allele counts to represent either a proportion out of a fixed number that reflects how many alleles are in the pool (i.e. how many chromosomes are being sequenced) or to the effective sample size n_{eff} (Kolaczkowski et al., 2011; Feder et al., 2012). Results from the re-analysis of the Orozco-terWengel et al., (2012) dataset suggest that n_{eff} is preferable.

Re-analysis of the Orozco-terWengel et al. (2012) data set also showed improvements in the consistency of the allele frequency difference between treatment lines across replicates in the top SNPs identified. The results were qualitatively similar to previously published analyses with peaks and troughs in the same genomic regions (Figure 3) although very few SNPs pass Bonferroni correction for multiple testing (Table 2.2). Furthermore, the large peak on chromosome 3 that is attributed to artefacts of higher coverage in Orozco-terWengel et al. (2012) is no longer visible (Figure 2.7).

In summary, the results presented here indicate that reliable identification of SNP alleles that occur at consistently different frequencies in different treatment lines across biological replicates of natural populations or experimental evolution lines requires two things. First, an appropriate statistical test needs to be chosen that does not confuse heterogeneity for a main effect. Two such tests, quasibinomial GLMs and linear models, are available and produce appropriate FPRs and TPRs, and also have other attractive properties that make them good tests to use. Second, it also emerges that variation in coverage across SNPs and replicates affects results in some circumstances. However, standardising coverage should be done with care because if the counts are too high this will create an artificially high level of confidence in overall effects, resulting in very low (effectively zero) p-values. The effective sample size procedure seems useful and is well grounded in theory (Kolacskowski et al., 2011; Feder et al., 2012). Finally, power (TPRs) seems to be related primarily to the number of replicates per treatment within the experiment although the strength of selection in comparison to the neutral divergence (F_{ST}) also plays a role .

This study uses a relatively simple simulation protocol which, while based on well established population genetic theory, makes some assumptions and has some limitations. For example, SNPs in the simulation are all independently sampled and are therefore effectively completely unlinked. It is a common issue with SNP-wise tests that linked SNPs will produce similar p-values simply because they will be affected by the same evolutionary processes. The extent to which this influences the interpretation of statistical results is not considered here but is an important aspect that could be included in simulations (e.g. Kessner & Novembre 2015; Baldwin-Brown et al., 2017). Similarly, the effect of selection was rather crudely modelled in this study and in reality the strength of selection will vary throughout the genome. In this context the high scoring SNPs in figures 2.9 and 2.10 may not be completely independent and similar allele frequency differences (and thus similar statistical results) will be observed for closely linked SNPs.

Throughout this study I have followed the convention that the more important loci to identify are those which diverge consistently across replicate treatment lines. It is commonly argued that such loci are those most likely to represent responses to divergent selection, because inconsistent divergence may be due to drift. However, it is

probably worth noting that evolutionary responses can often be opportunistic. Different SNPs segregating within genes or regulatory regions may provide alternate responses to similar selection pressures or some forms of selection (for example, parasite-host coevolution or sexual selection) may be particularly likely to cause inconsistent responses. Hence not all loci showing inconsistent responses in real datasets will be false positives.

2.5 Concluding Remarks

The increasing use of genomic methods in the comparative study of populations has introduced the need for novel analytical approaches. Many of these approaches are available in easy to implement software packages making them routine for researchers to use. However, this can sometimes lead to a "black boxing" effect where the tests are assumed to behave properly. In the study of stratified populations, such as experimental evolution studies, a recently proposed approach is to use CMH-tests to identify consistent allele frequency differences across replicates. However, this test is ill-suited for the task from first principles because genomic data frequently violate many assumptions. Sequenced reads of alleles are frequently used as independent counts of alleles which is a text-book case of pseudoreplication. Additionally, the very thing being tested (consistent differences) is a fundamental assumption of the test.

This study explored the behaviour of the CMH-test and some alternative (binomial GLMs, quasibinomial GLMs, and G-tests) to assess their performance in a hypothetical experimental evolution scenario. The main aim is to test the behaviour of these tests under the null hypothesis (neutral genetic drift). Additionally, a simple assessment of power is also undertaken. A population genetic simulation under neutral drift is performed which incorporates the unique allele sampling properties of a pool-seq experimental design. The results indicate that the CMH-test is indeed prone to very high rates of false positives, especially where coverage varies across the pooled samples. By contrast, quasibinomial GLMs and General Linear Models (LMs) peform much better and produce the characteristic uniform distribution of p-values expected under the null hypothesis. However, even these tests show an increased rate of false positives when coverage is allowed to vary across the pools.

The use of the popular CMH-test in the analysis of allele frequency differences in stratified populations is problematic. Additional complications arise from highly variable coverage in pool-seq experimental designs. This study identifies these problems and proposes some alternatives (quasibinomial GLMs and LMs) that behave well under the null hypothesis. These simulations highlight the importance of properly testing the performance of novel analytical approaches in a range of scenarios.

Chapter 3 The Genomic Response to Experimental Evolution Under Altered Mating Systems in *D. pseudoobscura*

Abstract

The interplay between sexual selection and sexual conflict is of great interest in evolutionary biology. Many studies have shown how traits are shaped by both sexual conflict and sexual selection, and how polyandry, the female re-mating rate, modulates the strength of selection. However, the genomics of sexual selection has been relatively unexplored. Experimental evolution is an excellent approach to studying the response to selection under novel environments. Combined with next-generation sequencing methods it can give great insight into genomic targets of selection.

In *Drosophila pseudoobscura* an experimental evolution study has been ongoing since 2002. The study was designed to test predictions about the response to altered mating systems in *D. pseudoobscura*, a naturally polyandrous fly. Females have been housed with either one or five males for the duration of their reproductive life. Because the female mating rate is an important part of mating systems these treatments are thought to eliminate sexual selection and sexual conflict or elevate it. Over the years studies have revealed a great deal of interaction between conflict between the sexes and sexual selection. Key traits in sexual selection, such as the male courtship song, have shown marked changes in response to these altered mating systems. Evidence also points to a greater capacity of males evolving under elevated rates of polyandry to coerce and manipulate females and achieve more matings in a competitive mating system. More recently, the genomic response to selection has started to be considered in this system.

This chapter focuses on pooled genome sequencing (pool-seq) data that was collected from each of four replicate lines of the experiment. The aim is to identify genomic markers (SNPs) which show consistent allele frequency differences between the experimental evolution treatments. Such consistent differences are prime candidates

for loci important to phenotypic changes observed in previous experiments. Additionally, if observed changes are a result of selection other genomic signatures should be observed, namely reductions in nucleotide diversity (π) and an allele frequency spectrum skewed toward an excess of rare variants (measured by e.g. Tajima's *D*). Finally, differences in the rates of polyandry makes predictions about the ratio of effective population size (N_e) on the X chromsome and autosomes. Differences in the mating system should therefore have an effect on relative levels of diversity and on population differentiation on the X chromsomes. These predictions about differences on X and autosomes have not previously been tested.

Author Contributions

The experimental evolution lines are maintained under enormous effort by Rhonda Snook (University of Sheffield) and her research group and sample collection was also performed by them. Extraction of genomic DNA and all subsequent analyses presented here are my own work.

3.1 Introduction

3.1.1. Sexual Selection and Sexual Conflict

Charles Darwin (1859; 1871), described the process of sexual selection, the competition among members of one sex for the reproductive resource of the other sex, as being distinct from natural selection, selection for viability and fecundity. Since then, these forces and how they interact to drive the evolution and diversification of some of the most extreme traits in nature have captivated biologists. A great body of theoretical models and empirical evidence has accumulated to clarify how sexual selection can drive the exaggeration and coevolution of traits (e.g. Andersson 1994) and though views of where to draw the line between sexual selection and natural selection are now more nuanced the distinction remains in use (Shuker 2014). More recently, sexual conflict, "conflict between the evolutionary interests of individuals of the two sexes" (Parker 1979) between the two sexes have begun to receive more attention (Trivers 1972; Parker 1979; Arnqvist & Rowe 2005). Trivers (1972) described situations where conflict occurs between the sexes over the provisioning of parental care. Later, Parker (1979) developed a genetic model that showed that a trait that increases male mating

success but comes at a cost to fecundity in females can spread through a population (Parker 1979). Parker also pointed to examples of such traits in the aggressive attempts to mate with receptive females in yellow dung flies (*Scatophaga stercoraria*), and in the mate guarding bahavior of ground beetles (*Cicindela maritima*) (Parker 1979). Finally, he also realised that it would be possible for the sexes to end up in an arms race or "evolutionary chases" which has since developed into a theory commonly referred to as Sexually Antagonistic Coevolution (SAC) (Parker 1979; Holland & Rice 1998; Perry & Rowe 2014).

Over the years these models of sexual selection and sexual conflict have been further developed and empirical work has found some support for sexual conflict being a strong diversifying force in nature (Arnqvist & Rowe 2005; Tregenza et al., 2006; Perry and Rowe 2014). Sexual conflict is typically considered not as a distinct force but as the outcome of "multiple components of natural and/or sexual selection [that] favour certain trait values in males and other trait values in females" (Shuker 2014, pp 25**)**. Sexual conflict interacts very tightly with sexual selection, in fact "...sexual selection is an inevitable facet of sexually antagonistic coevolution." (Arnqvist & Rowe 2005, pp 35). This is partly because sexual conflict is likely to be present in every mating system except strict monogamy since a major source of conflict can be the mating rate itself (Arnqvist & Rowe 2005; Arnqvist and Nilsson 2000). Males typically favour a much higher rate than females but at the same time males benefit from low re-mating rates among females with which they have mated (Parker 1979). Despite this conflict, polyandry is widespread (Pizzari & Wedell 2013, Taylor et al., 2014; Snook 2014). The mating system of an organism, polyandrous ones in particular, plays a significant role in determining the strength of both sexual selection and conflict (Pizzari and Wedell 2013), so disentangling the effects of sexual conflict and sexual selection is difficult.

There are many reasons why polyandry might initally be favoured in different contexts and be maintained due to sexual or natural selection (Snook 2014). Polyandry might be an adaptive trait insofar as it provides direct or indirect benefits to the female. However, polyandry can also be entirely non-adaptive and arise from conflict over the mating rates in males and females (Snook 2014). The causes that promote the origins of polyandry can be distinct from those maintaining it (Snook 2014). Polyandry introduces new opportunities for sexual selection and conflicts between males and females in

mating decisions. For example, polyandry is expected to increase both pre- and post-copulatory sexual selection via female choice and sperm competition (Snook 2014). Examples of traits which are thought to evolve *via* sperm competition or mate choice resulting from polyandry, are adapations in sperm morphology (Snook et al., 2005), sperm number, ejaculate size and quality (Evans & Simmons 2008), and mate guarding bahaviors (e.g. Burdfield-Steele & Shuker 2014; Parker & Vahed 2009), among others. Meanwhile, polyandry often carries costs to females (Arnqvist & Nilsson 2000) resulting in further conflicts over the optimal mating rate between the sexes. Likewise, polyandry can lead to selection for traits that increase female fitness in the face of male adaptations to intense sexual selection and conflict. In particular, the bewildering diversity of seminal fluid proteins in various insects which influence female mating and ovulation rates are a classic example of traits thought to be adapatations in the face of conflicts over mating rates (e.g. Chapman 2008). Thus, polyandry, sexual conflict and sexual selection are closely linked in mating system evolution.

A vast literature exists on how different male and female traits interact to determine the outcomes of the intra- and inter-sexual competition for reproductive resources (see e.g. Andersson 1994; Arnqvist & Rowe 2005; Perry & Rowe 2014 for reviews), and while relatively little is known about the genetic basis of these traits or the genomic targets of selection (with the exception, perhaps, of model systems like *Drosophila*) the field is rapidly adopting genomic methods to give new insights on the effects of mating systems and sexual selection on genomes (Ritchie and Butlin 2014; Wilkinson et al 2015).

3.1.2 The Genomics of Sexual Selection and Sexual Conflict

A genomic perspective is worth pursuing for the advances it will bring to our understanding of how adaptations can arise and are selected for and built at the genomic level by various forces of selection. For example, understanding how sexually dimorphic traits are built from a genome shared between males and females is an important goal. It is likely that most sexual dimorphisms are produced by differences in gene expression (Williams & Carroll 2009; Wilkinson et al., 2015). If this is the case then the loci involved in the divergence of such dimorphic characters between species must include regulatory regions as well as coding sequences (Wilkinson et al., 2015).

Studies that alter mating systems (and presumably levels of sexual selection and conflict) frequently report changes to gene expression patterns (Hollis et al., 2012, Innocenti et al., 2013; Gerrard et al., 2013; Immonen et al., 2014; Perry et al., 2014; Hollis et al., 2016a). This suggests that regulatory regions may indeed be under strong selection in these systems. Though the treatments are not identical in each case and gene expression assays differ across studies some interesting patterns are clear. Expression differences are often in genes that are initially sex-biased (Hollis et al., 2012; Hollis et al., 2016a), localise to reproductive tissues (Innocenti et al., 2013; Immonen et al., 2014), and involved in the post-mating physiological manipulation of female egg-laying and re-mating rates (Perry et al., 2014; Hollis et al., 2016a). Such categories are ones that might be expected under changes to sexual selection or conflict.

The sex chromosomes, on account of their differences in zygosity in males and females, undergo different patterns of evolution from autosomes (Vicoso & Charlesworth 2006). For example, in XY systems, where males are hemizygous for the X chromosome, selection is expected to be more efficient on the X chromosome because recessive mutations are always exposed in males (Vicoso & Charlesworth 2006). However, mating systems, sexual selection, and sexual conflict also have the potential to affect the rates of evolution and the effective populations sizes of X and Y chromosomes in a population. All else being equal, with an equal number of males and females, random variation among males in the number of offspring produced should result in a ratio of effective population size (N_e) on the X chromosome and the autosomes of 0.75 (Ellegren 2009; Mank et al., 2010; Corl & Ellegren 2012). This also means that diversity should always be lower on the X (or Z) chromosomes than on autotosomes (Ellegren 2009). The ratio of N_e will affect the rate of evolution on the X chromosomes. Deviations from neutral expectation such as differences in the mating systems, demographic forces, or changes in the variance in male mating success owing to selection are expected to affect sex chromosome evolution by altering N_e of the sex chromosomes relative to the autosomes (Mank et al., 2010; Corl & Ellegren 2012). For example, higher rates of polyandry, and the associated reduction in variance of male mating success, (in XY systems) should result in a lower X:A ratio of N_e which in turn, is expected to reduce the X:A ratio of diversity (Charlesworth 2001; Ellegren 2009; Corl & Ellegren 2012).

Although deviations form the 3:4 ratio are frequently seen in wild populations the cause is not clear. In humans (Arabiza et al., 2014), *D. melanogaster, D. simulans* and some birds (Ellegren 2009), ratios of diversity are variable across populations in a manner that is broadly consistent with historical migration and associated demographic changes. Others have argued that deviations from the 3:4 ratio could be due to differences in recombination or mutation rates between the X and autosomes (Vicoso & Charlesworth 2006). Meanwhile, in a comparative study of matched sister species of shorebirds that differ in their mating systems, more polygynous species have a reduced diversity on the Z chromosome compared to autosomes as expected if sexual selection is a potent force in producing relative patterns of diversity on the sex chromosomes and autosomes (Corl & Ellegren 2012). A comparison of closely related flycatchers (*Ficedula hypoleuca* and *F. albicolis*) found that the Z:A ratio is much lower than expected from a neutral model even under extreme operational sex-ratio biases, indicating selection (Borge et al., 2005). However, demographic effects, such as population contractions and expansions could also be contributing to these patterns. Population genomic analysis of the same species pair found higher overall levels of differentiation and fewer shared polymorphisms on the Z chromosome than autosomes, as well as a more uniform LD (Ellegren et al., 2012). These results are consistent with both a greater role for selection and reduced gene flow on the Z chromsome compared to autosomes. Another comparative study on the relative rate of evolution on Z and autosomes finds a similar pattern (Wright et al., 2015). However, faster evolution on Z chromosomes in polygynous species seems to be mainly due to genetic drift as a consequence of the reduced N_e (Wright et al., 2015). Meanwhile, in *Drosophila* (XY system) there is some evidence for more efficient selection on the X chromosome because deleterious recessive alleles are always "visible" to selection in males (Vicoso & Charlesworth 2006; Langley et al., 2012).

Clearly, more comparative studies are needed in systems that differ in their mating systems while controlling for other factors. In particular, comparative studies in XY systems and the effects of polyandry, have not been investigated. A corollary of lower diversity on the X chromosomes is that the ratio of F_{ST} on X chromosomes should be higher between populations (Ellegren 2009). Thus if polyandry has a large effect on the strength of sexual selection and sexual conflict it should drive patterns of diversity

and differentiation on X chromosomes.

In summary, a few points can be highlighted on the state of understanding of natural and sexual selection, and sexual conflict. First, the interplay of sexual conflict and sexual selection in the evolution and diversification to various traits is likely to be significant. Secondly, the mating system, characterised in large part by the rate of female re-mating (polyandry), plays a major role in determining the strength of sexual selection and conflict. Third, although evidence points to a large role for transcriptional regulation, the genomics of sexual selection and sexual conflict remains relatively understudied. Finally, other theoretical predictions about the effects that different mating systems have on the effective population size and selection on sex chromosomes are also relatively understudied. Thus, studies that aim to understand the effects of mating system difference, as well as the interplay of sexual conflict and sexual selection on the genome and the targets of selection are needed.

3.1.3 Experimental Evolution

An excellent way to study adaptation to different environments and test predictions from theory is to set up experimental evolution studies (Arnqvist & Rowe 2005; Kawecki et al., 2012; Schlötterer et al., 2015). These allow a researcher to manipulate characteristics of an organism's environment (including the social environment) while controlling other factors in order to observe evolutionary responses. The falling costs of Next Generation Sequencing (NGS) technologies have opened up new possibilities to identify the genetic variants that differ between populations. Additionally, the sequencing of pools of individuals (pool-seq) rather than individuals allows larger sample sizes and replicated experimental evolution lines at reasonable costs (Schlötterer et al., 2014). Experimental evolution studies in combination with population genomic methods, typically called Evolve and Resequence (E&R) studies, are growing in number to answer a wide range of questions in evolutionary biology. Such studies have great potential to answer questions about the genomic targets of selection and how population differences arise (Kawecki et al., 2012; Schlötterer et al., 2015).

Experimental evolution studies have been used to study responses to changes in sexal selection and sexual conflict in various organisms because the method naturally

lends itself to altering the mating system by manipulating the ratio of males to females in different treatments (Arnqvist & Rowe 2005). A classic example is the study by Holland and Rice (1999). Female *D. melanogaster* were kept under conditions of forced monogamy for 47 generations by allowing a female to mate only with one male. Sexual conflict theory predicts that conflict over the re-mating rate should result in harmful males and therefore the removal of this conflict should select for less harmful males and conversely females that are less resistant to male harm (Holland & Rice 1999). Indeed, females from the final generation of the monogamy treatment had reduced longevity compared to ancestral (control) females when exposed to ancestral males which courted more often, indicating ancestrally harmful male matings (Holland & Rice 1999). Other experimental evolution studies under altered mating systems have been performed in dung flies (*S. stercoraria*; Hosken & Ward 2001; Hosken et al., 2001), robber flies (*Sepsis cynipsea*; Martin & Hosken 2003), fruit flies (*D. melanogaster*; Holland and Rice 1999; Hollis et al., 2012, 2016a, 2016b, Innocenti et al., 2013; Gerrard et al., 2013; Perry et al., 2014; *D. pseudoobscura;* Crudgington et al., 2005), and in free living flatworms (*Macrostomum lignano*; Janicke et al., 2016).

Although these studies differ in the specifics of their designs they all vary the general levels of sexual selection and conflict by controlling the opportunity for re-mating and aggressive courting. These studies all find evidence that some traits respond rapidly to selection from altered mating systems and give evidence for both sexual conflict and sexual selection in these systems. For example, it is a common observation that females evolving under higher levels of polyandry become adapted to resist or better cope (measured as post-mating survival) with the costs associated with continuous courting by and costly matings with males (Holland and Rice 1999; Martin & Hosken 2003; Innocenti et al., 2013). The corresponding observation that monogamy relaxed selection for, or direct selection against, resistance and coping is also observed (e.g. Hollis et al., 2016a).

3.1.4 Experimental Evolution in D. pseudoobscura

Since 2002, an experimental evolution experiment has been underway using *D. pseudoobscura,* a naturally polyandrous fruitfly, to address how populations adapt to changes in mating systems via sexual selection and sexual conflict (Crudgington et al.,

2005). A full description of the experimental treatments is given in Crudgington et al., (2005). The treatment regimes alter the mating system by housing females with either one male (M; enforced monogamy) or five males (E; elevated polyandry) (Crudgington et al., 2005, 2010; Bacigalupe et al., 2007). Enforced monogamy is assumed to eliminate both sexual selection sexual conflict while elevated polyandry maintains or increases both sexual selection and sexual conflict (Bacigalupe et al., 2007; Crudgington et al., 2005, 2009). Phenotypic changes were observed across replicated experimental evolution treatments from the early stages of the experiment. After 20-31 generations, mating trials of male and female combinations from different experimental treatments showed that there was a significant effect of the number of mates on female fecundity and the number of eggs hatched, but not of male evolutionary history (Crudgington et al., 2005). Meanwhile E treatment females were more fecund when mated to ancestral males suggesting that they are better adapted to resist costs of mating with competitive males. Finally the evolutionary history of males influenced their ability to manipulate the female willingness to re-mate but not in the ability to coerce already-mated females to re-mate (Crudgington et al., 2005).

Later work investigated how variation in sexual selection would impact the evoluton of male traits, for example courtship song traits are an important source of variation in male mating success under different mating systems (Snook et al., 2005). They found that E males had shorter “inter-pulse-intervals” and were faster to begin producing song (Snook et al., 2005). Another analysis at 110 generations replicated these results (Debelle 2013). These studies also found that variation in song characters among replicates of E treatments was greater (Snook et al., 2005). Sexual conflict theory predicts that coevolution of reproductive traits could produce reproductive isolation between allopatric populations. To test these predictions with flies from generations 48-52, Bacigalupe et al., (2007) performed “sympatric” and “allopatric” crosses of flies from different treatment lines. “Sympatry” in this case was a cross in which both the male and female were from the same replicated treatment line while in “allopatric” crosses males and females were from different treatment lines. No differences in mating speed, copulation duration, or the number of mating paris were observed nor were any hybrids sterile or less viable (Bacigalupe et al., 2007). A similar study found no difference in the latency to mate or the mating outcome (E males always

win in competitions with M males) in different crosses of males and females (Debelle et al., 2016)

High levels of post-copulatory sexual selection via sperm competition and sexual conflict both predict responses in male behaviour. Males should mate more often and/or invest in manipulative ejaculate components and sperm (Crudgington et al., 2009). Several experiments between generations 42 and 78 suggested that evolutionary history did not explain variation in sperm morphology traits or testis mass, but E males had larger accessory glands (despite no differences in mean body size among treatments at this stage). E males were also able to mate with more females sequentially (greater mating capacity; Crudgington et al., 2009). However, heteromorphic sperm, a characteristic "sterile class" of sperm within *D. pseudoobscura,* did not change in morphology or number. This suggests that heteromorphic sperm traits do not seem to be driven by sexual selection in *D. pseudoobscura* (Crudgington et al., 2009).

As highlighted above, polyandry is also predicted to generate conflict and select for male traits that harm females to prevent re-mating or coerce already-mated females to re-mate. To test this prediction, Crudgington et al., (2010) housed males and females from generations 54-55 of different treatments in mating trials. Reproductive output of M females was lower if housed with E males than when housed with M males. Meanwhile, M females confined with E males produced a greater proportion of their progeny in the first 7 days than M females with M males. In contrast to earlier results, lifetime reproductive output of females housed with E males produced fewer offspring. However, female survival was not related to the evolutionary history of male. Finally, E males courted more than M males (Crudgington et al., 2010). These results suggest that sexual conflict does promote harmful males.

Since earlier work showed changes in song characters (Snook et al., 2005), it is also possible that female preferences might change in response to sexual selection. Females from each of generations 131 through 135 of the experimental evolution treatments prefer the song type of co-evolved males (Debelle et al., 2014). This preference also seems to extend to more extreme versions of songs suggesting a change in some intrinsic bias among the females for particular song characteristics. Curiously, song preferences also changed in M females despite the prediction that female choosiness in these lines would be very costly as they only mate with one random male

(Debelle et al., 2014). These changes in preferences do not seem to lead to assortative mating however, since E males have a higher probability of mating than M males in competitive trials with females from both treatments (Debelle et al., 2016).

While several experimental evolution studies have considered the genomic response to selection under novel environments (Burke et al., 2010; Orozco-terWengel et al., 2012; Barrick & Lenski 2013; Martins et al., 2014) none have addressed the genomic basis of traits under sexual selection or conflict except in terms of changes to gene expression (Hollis et al., 2012, Innocenti et al., 2013; Gerrard et al., 2013; Immonen et al., 2014; Perry et al., 2014; Hollis et al., 2016a, see above). As in other systems, when female transcriptomes are compared, E females show an increase in expression of genes normally enriched in ovaries (Immonen et al., 2014).

Here I investigate patterns of genomic change in response to experimental manipulation of sexual selection and sexual conflict using a pooled sequencing approach. The aim is to uncover genetic variants (SNPs) that show consistent allele frequency differences between the E and M lines. Such differences are most likely to be the result of selection during the experiment and may therefore identify the targets of selection within the genome in response to altered mating systems. I make use of the methods that I developed and evaluated in Chapter 2 to perform these analyses. These regions might also be expected to show patterns of diversity consistent with selective sweeps in response to changes in selection. Additionally, we might expect transcription factor binding motifs to be enriched in regions showing consistent allele frequency changes. Finally, if differences in mating systems and sexual selection has a big effect on X chromosomes, then patterns of F_{ST} and diversity on the X chromosome should be greater than expected from differences in N_e between autosome and the X chromosome alone.

3.2 Methods

3.2.1 Sequencing and Mapping

Whole-genome sequencing was carried out by the NBAF sequencing facility at the Center for Genomic Research (CGR) within the University of Liverpool. Samples were sequenced using a "pool-seq" approach (Schlötterer et al., 2014). For each

experimental evolution line, 40 females were pooled from generations 164 for replicate 1, 163 for replicate 2, 162 for replicate 3, and generation 160 for replicate 4 of the experiment and DNA was extracted using a standard phenol-chloroform extraction protocol (Appendix A). 8 libraries were run on a single Illumina HiSeq lane and sequenced to ~40x coverage.

Quality control by trimming and filtering low quality reads was performed using Trimmomatic v. 0.32 (Bolger et al., 2014). Reads were clipped if the base quality is < 20 and reads shorter than 20 bp were discarded. Reads were mapped to the *D. pseudoobscura* reference genome (release 3.1, Februrary 2013), obtained from FlyBase (dos Santos et al., 2014), using BWA mem (Li 2013**;** Li & Durbin, 2009). Following "best practice" recommendations for pool-seq studies (Schlötterer et al., 2014) duplicate reads (representing identical PCR products from the same individual) were removed using samtools v. 1.2 (Li et al., 2009) and re-alignment around indels was carried out in GATK v. 3.3 (McKenna et al., 2010; DePristo et al., 2011). Finally, Bedtools v. 2.22.1 (Quinlan & Hall, 2010) was used to calculate various genome-wide statistics like coverage throughout the genome. SNPs and allele frequencies were called with samtools v. 1.2 (Li et al., 2009) and PoPoolation2 v1.201 (Kofler et al., 2011). A number of quality filtering criteria were used to identify SNPs. Only biallelic SNPs were considered, if a SNP had more than 16 reads for a third allele the SNP is considered multiallelic and discarded. Additionally, a minimum and maximum coverage threshold of 17x and 49x (the 10^{th} and 90^{th} quantiles of the aggregate coverage distribution respectively respectively; figure 3.1) are applied. If any sample falls outside of these thresholds for a given SNP that SNP is discarded.

3.2.2 Patterns of Genome-wide Variation

Various measures of genome-wide variation are calculated using the PoPoolation (π and Tajima's D, Kofler et al., 2011**)** and PoPoolation2 (pairwise F_{ST}*;* Kofler et al., 2011) packages. π and Tajima's D are estimated for non-overlapping windows of 10 kb within each treatment line. SNPs are only counted if alleles have a minimum read count of 8, if the coverage in within each replicate is > 17 and < 49. For F_{ST}, overlapping windows of 50kb (with a 5kb overlap) were chosen and F_{ST} calculated for each window on a pairwise basis using the same read count and coverage

threshholds as above. Similarly, allele frequency differences are first calculated for each SNP using PoPoolation2 and then summarised as averages in non-overlapping windows of 50kb across the genome. From the allele frequency differences it is possible to infer d_{XY} between treatment lines of the same replicate. Here d_{XY} is simply estimated as the proportion of fixed differences between two treatments within a given window as follows:

$$nr.\ fixed\ differences\ /\ window\ length$$

For F_{ST}, d_{XY} and allele frequency differences experimental evolution lines were paired within replicates. Within a replicate pair of E and M treatment lines, F_{ST} and d_{XY} was calculated for these pairs. This pairing is justified on the basis of the initially staggered establishement of these experimental evolution lines (see *3.1 Introduction*).

The expected F_{ST} on X chromosomes (F_X) was calculated as in Machado et al., (2016) using the equations of Ramachandran et al., (2005). F_X is given by:

$$F_X = 1 - \left\{ \frac{\left(9(z+1)(1-F_A)\right)}{\left(8(2z+1) - (1-F_A)(7z-1)\right)} \right\}$$

where, z is the ratio of the number of breeding males to females and F_A is the observed F_{ST} on autosomes. F_X was calculated for z = 1 and 5. In the pairwise analyses between E and M treatments from each replicate (see above), F_{ST} was first averaged across all windows on a chromosome to give a chromosome-wide F_{ST} value. Then an average F_{ST} across the main chromosomes was calculated and converted to F_X. A bootstrapping approach was used to obtain a distribution of F_X. For each bootstrap iteration a sample of windows equal to the total number of windows across all chromosomes was sampled, with replacement, from the set of all windows. Then the average F_{ST} was calculated across all windows and then converted to F_X. This sampling was repeated 1,000 times to obtain a distribution of F_X for each replicate.

In a separate analysis, F_{ST} was calculated for each pair of replicate lines within a treatment group. That is, for each treatment (E and M), all pairwise comparisons across

replicate lines were performed. Then for each replicate line, the average F_{ST} across all pairwise comparisons, within a window, which included that line was taken as an overall F_{ST} for that line. A bootstrap sampling was performed as above to give a distribution of F_X.

3.2.3 Identifying Candidate SNPs

A GLM with a quasibinomial error distribution (QB-GLM) was fitted to the allele counts at each SNP to identify SNP alleles that are found at consistently different frequencies between treatment lines, across replicates. The structure of this model is:

$$y \sim treatment + e$$

Where y is the allele frequency expressed as a proportion, *treatment* is a factor giving the experimental evolution treatment line of origin (E or M), and e is an error term with a quasibinomial distribution.

Although a biologically meaninful pairing of treatment lines exists in this case, an unpaired QB-GLM is applied because there are not enough degrees of freedom to test for a paired effect in these data. Thus, replicate is not included in the model. This procedure fits a QB-GLM with experimental evolution treatment as a fixed effect to the SNP allele frequencies from each treatment line. Spurious results can occur if there are too many zero counts in the data (see Chapter 2). To avoid this, a count of one is added to all allele counts if any allele count of zero count are detected in any population. Because QB-GLMs behave well under a null hypothesis (see Chapter 2) it is possible to take advantage of standard corrections for multiple test correction. A python script for fitting the QB-GLM to these data is available from the GitHub onlide code repository (https://github.com/RAWWiberg/poolFreqDiff). The False Discovery Rate (FDR) is controlled by q-values (calculated from the distribution of p-values using the "qvalues" package in R (Dabney & Storey 2013)) (Storey 2002, Storey & Tibshirani 2003). The SNPs with a q-value < 0.05 are taken as the "candidate" SNPs that show a consistent evolutionary response to experimental change in the strength of sexual selection and sexual conflict. In chapter 2 I recommend scaling read counts to counter the problem of pseudoreplication in. However, here the number of SNPs called as significant is very

similar whether this procedure is used or not. All of the candidate SNPs identified using raw counts are recovered after scaling to the effective sample size (n_{eff}; Kolaczkowski et al., 2011; Feder et al., 2012). Scaling to n_{eff} adds an additional 103 significant SNPs meaning that there are fewer significant SNPs when using the raw counts. Therefore candidate SNPs from raw counts are used for downstream analyses.

3.2.4 Functional Analysis

GO term enrichment analysis is carried out using Gowinda v. 1.12 (Kofler & Schlötterer, 2012) using 1 million simulated permutations to obtain the empirical distribution of gene numbers for each GO term (Kofler & Schlötterer, 2012). Within GOwinda, SNPs that occur within 1 kb or 1 Mb up or downstream of a gene are considered as "genic" and all associated genes are kept in the analysis. These physical distances are justified on the basis that enhancer regions can occur up to 1 Mb up- or downstream from a target gene (e.g. Maston et al., 2006; Pennachio et al., 2013). Recent reports of variant sites in enhancer regions influencing colour patterns in *D. guttifera* report the causal loci at distances of ~5kb fron the relevant genes (Werner et al., 2010)**.** Similarly, recent reports from sticklebacks show that the effect of locus *Pitx1* on adaptive pelvis reductions among freshwater populations is controlled by an enhancer region ~23kb upstream of *Pitx1* (Chan et al., 2010). Gene ontology (GO) terms are not available for *D. pseudoobscura* so only genes which have orthologs in *D. melanogaster* are considered (11,622 loci). Known gene duplications between *D. melanogaster* and *D. pseudoobscura* exist, thus *D. melanogaster* genes that appear as orthologous with multiple *D. pseudoobscura* genes are separately labelled and kept in the analysis (870 loci). GO terms for *D. melanogaster* genes are downloaded via FuncAssociate (v. 2) (Berriz et al., 2009) which queries and combines information from several databases including Gene Ontology and Ensembl (Berriz et al., 2009). Additionally, the closest gene to each SNP is identified with bedtools (v. 2.17.0; Quinlan & Hall 2011) closestBed (keeping any potential ties).

Previous studies have shown differences in gene expression in virgin females from E and M treatments in these same experimental evolution lines (Immonen et al., 2014). More recently, a transcriptome study was carried out contrasting transcription patterns of genes in different tissues, sexes, and individuals of differing mating status

(virgin or courted) in the same experimental evolution treatment lines (Veltsos et al., *in prep*). With these data it is possible to ascertain whether genes near the candidate SNPs (candidate genes) were over-represented among those genes identified as being differentially expressed (DE) in Immonen et al., (2014) or Veltsos et al,. (*in prep*). A rough estimate of the significance of any enrichment can be obtained by picking a random sample of *n* genes from the *D. pseudoobscura* genome, where *n* is the number of candidate genes associated with the candidate SNPs, and asking how many occur in the list of genes that are DE in the Immonen et al., (2014) or Veltsos et al., (*in prep*) datasets. This procedure is repeated 10,000 times and the proportion of times the random values are greater than or equal to the observed value for the candidate genes is taken as an empirical p-value.

Because gene expression changes are observed in these experimental evolution lines, it is possible that regulatory regions are under strong selection. To investigate whether regulatory regions are an important target of selection, a transcription factor (TF) motif enrichment analysis is performed. The region comprising 30 bp in either direction around the positions of the candidate SNPs within 1 kb and within 1 Mb upstream of a gene are extracted from the reference genome. The motivation is that a causal SNP might alter a critical part of a motif and thereby change TF binding affinity. A region of 30bp is sufficiently wide to contain a full motif but not so wide as to exclude the focal SNP from these motifs. Enrichment of TF binding site motifs was performed using the AME tool from the MEME v4.10.2 package (Bailey et al., 2009; McLeay & Bailey, 2010). Enrichment of motifs in the candidate SNP sets is compared to the background of all discovered SNPs. Motifs for two transcription factors; *doublesex* (dsx) and *fruitless* (fru), in particular, are tested for enrichment in the region around the candidate SNPs. Position weight matrices for these motifs are available for *dsx* from Clough *et al.* (2014) as well as the Fly Factor Survey (Zhu et al., 2011) and for *fru* from the Fly Factor Survey. The FIMO tool in the MEME suite also finds any TF binding motifs present within the query sequence and computes an assignment confidence.

The online software DroPHEA (Weng & Liao, 2011) is used to conduct phenotypic enrichment analysis. This procedure is similar to GO enrichment but instead of genes belonging to GO groups they belong to groups defined on the basis of the

phenotypic effects of mutations at these genes. Phenotypic effects are collated from the online database FlyBase (data release date: January 2016) (Weng & Liao, 2011; dos Santos et al., 2014). DroPHEA allows the user to specify classes of phenotypic effects into user-defined phenotypes. The phenotypic classes "courtship bahavior defective" (FBcv:0000399) and "mating rhythm defective" (Fbcv:0000401) were combined into one phenotype group and the phenotypic class "stress response defective" (Fbcv:0000408) was also tested for enrichment. The genes within 1 kb or 1 Mb of the candidate SNPs were tested for enrichment of the combined classes compared to the rest of the genome using *D. melanogaster* gene IDs by a Fisher's exact test.

3.3 Results

3.3.1 Sequencing and Mapping

The number of mapped reads, proportion of all reads mapped, and average coverage is shown in table 3.1. These results suggest a fairly even coverage across all samples. The distribution of coverage per base throughout the genome is shown in figure 3.1. The fourth and X chromosomes are not fully assembled in the reference genome, so the individual segments of these chromosomes are kept separate in the figures below.

Table 3.1. Summary statistics of sequencing, quality filtering and mapping steps for each sample. Given are the number of reads, the proportion that are mapped, the proportion that are properly paired (i.e. the forward and reverse reads align in the correct orientation to the same chromosome), and the mean coverage.

Replicate	***N reads (m)***	***Mapped (paired) (%)***	***Coverage***
E1	~60.17	100 (98.08)	42.04x
E2	~65.92	100 (98.04)	45.73x
E3	~64.90	100 (98.06)	45.32x
E4	~52.77	100 (98.05)	37.05x
M1	~58.55	100 (97.99)	41.03x
M2	~56.08	100 (97.99)	39.53x
M3	~46.80	100 (98.08)	33.19x
M4	~52.31	100 (98.08)	37.02x

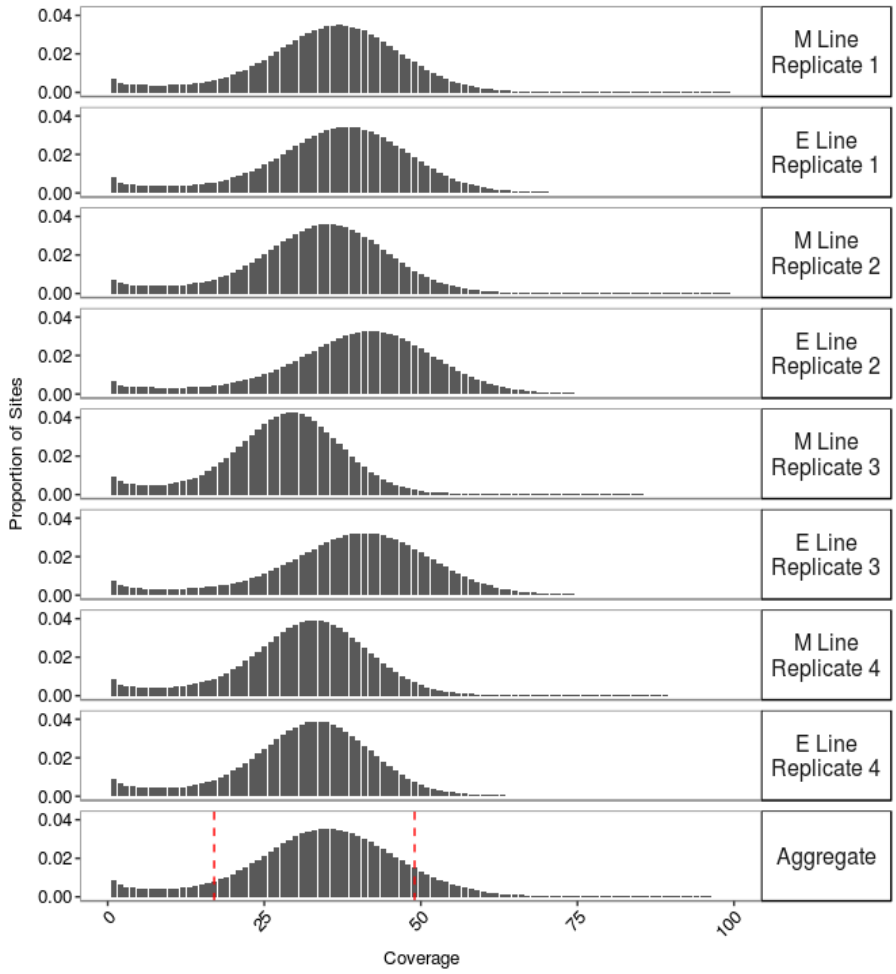


Figure 3.1. The distribution of coverage for each sequenced sample. The red vertical lines in the bottom panel indicate the 10th (17x) and 90th (49x) quantiles of the aggregate distribution. See table 3.1 for summary statistics from the mapping steps.

3.3.2 Patterns of Genome-wide Variation

Mean pairwise F_{ST} between E and M lines across replicates was lower on autosomes (0.46+/-0.003) than on the X chromosomes (0.64 +/- 0.003). Observed F_{ST} is significantly greater on the X chromosome than expected simply from differences in N_e (figure 3.2). This is true regardless of the ratio used for z (the ratio of breeding males to females). Additionally, X chromosome to autosome ratios of F_{ST} in pairwise comparisons among replicate lines are significantly higher in E lines (mean ± SD: 1.53 ± 0.08) than in M lines (1.30 ± 0.06). Among pairwise comparisons within E and M

lines, F_{ST} is always greater than expected from differences in N_e on the X chromosome (figure 3.3). These differences are not driven by differences on the autosomes where F_{ST} is the same across comparisons in E (0.37 ± 0.03) and M (0.39 ± 0.01) lines. Pairwise F_{ST} between E and M treatments, within replicates, throughout the genome are shown in figure 3.4. F_{ST} is generally quite high throughout the genome with some regions (e.g. the distal end of chromosome three) showing "peaks" of F_{ST} (figure 3.4). Figure 3.5 shows a measure of the proportion of sites that are fixed between E and M lines within non-overlapping 50kb windows throughout the genome. d_{XY} is also variable throughout the genome with a higher proportion of substitutions between treatment lines on the X chromosomes (figure 3.5).

Figure 3.6 shows the level of nueleotide diversity (π) within each replicated set of treatment lines. Overall π is greater in M lines than in E lines. This is true for autosomes (mean ± 2·SE for E: 0.0058 ± 0.001, M: 0.0060 ± 0.0005) and X chromosomes (mean ± 2·SE for E: 0.0020 ± 0.001, M: 0.0027 ± 0.001). Only in replicate 2 is π greater in E than in M lines, though the differences are not significant even if replicate 2 is excluded. The overall X:A ratio of π is much lower within E lines (mean ±2·SE for E: 0.34 ± 0.16, M: 0.45 ± 0.14). Again, these differences are not significant even if replicate 2 is excluded. As with F_{ST}, the distal end of chromosome three stands out as a region of low nucleotide diversity (figure 3.6).

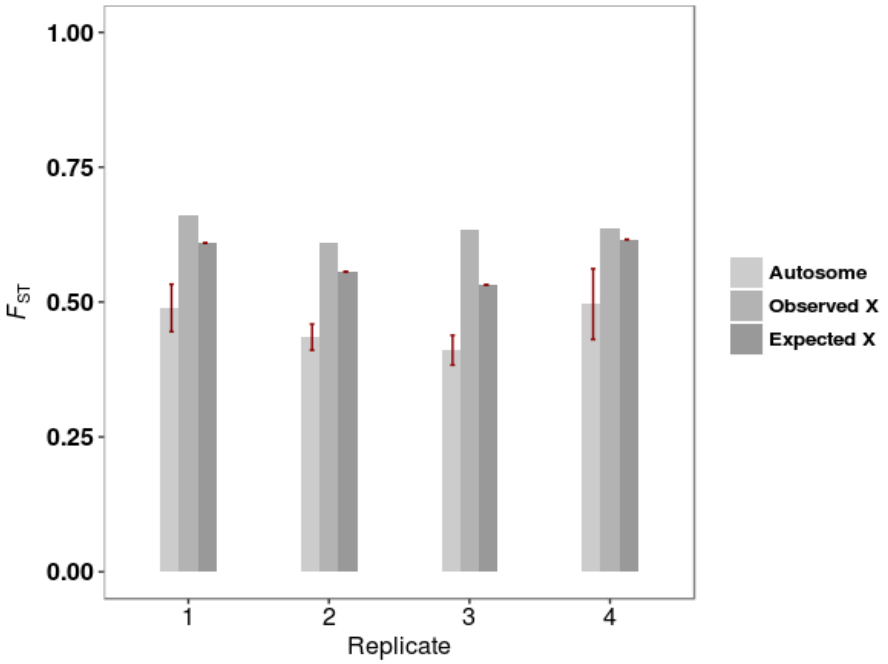


Figure 3.2. Autosomal and X chromosome F_{ST} between pairs of E and M lines. Observed values are the chromosome-wide mean F_{ST}. Autosomal error bars represent the standard deviation across chromosomes. The ratio of males to females (z) is assumed to be 5. Results for z = 1 are not shown. Error bars for the expected F_{ST} represent 2 * the standard error of a bootstrap distribution of values (see *3.2 Methods*).

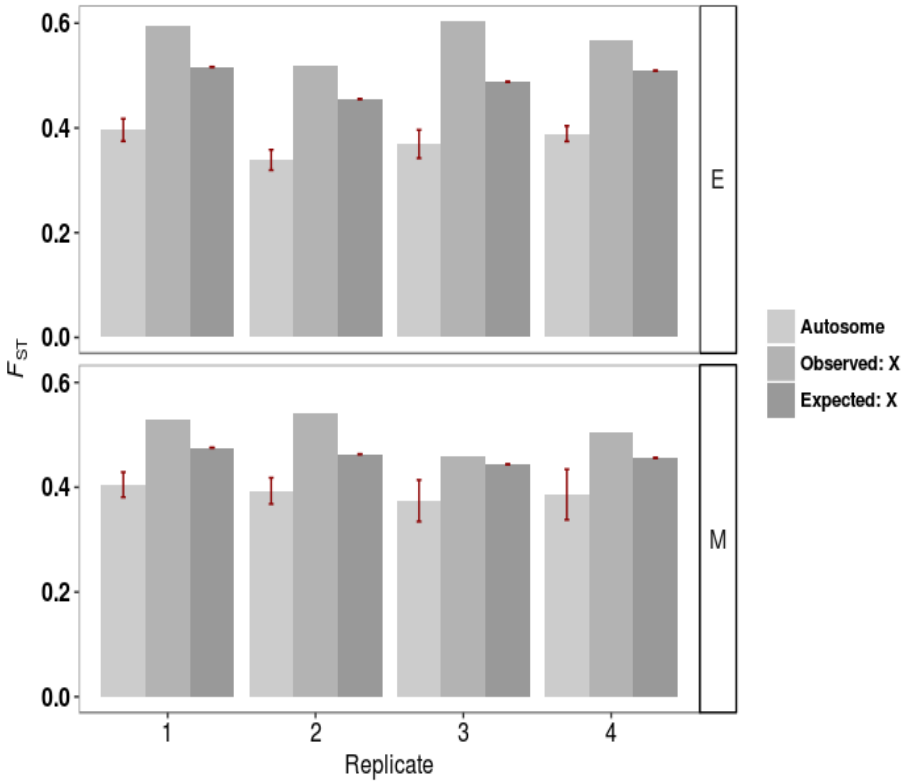


Figure 3.3. Autosomal and X chromosome F_{ST} for each replicate line within the E and M treatments. Error bars on the autosomes are the standard deviation across the main chromosomes. Error bars for the expected F_{ST} represent 2 * the standard error of a bootstrap distribution of values (see *3.2 Methods*). The ratio of males to females (z) is assumed to be 5 for E lines and 1 for M lines.

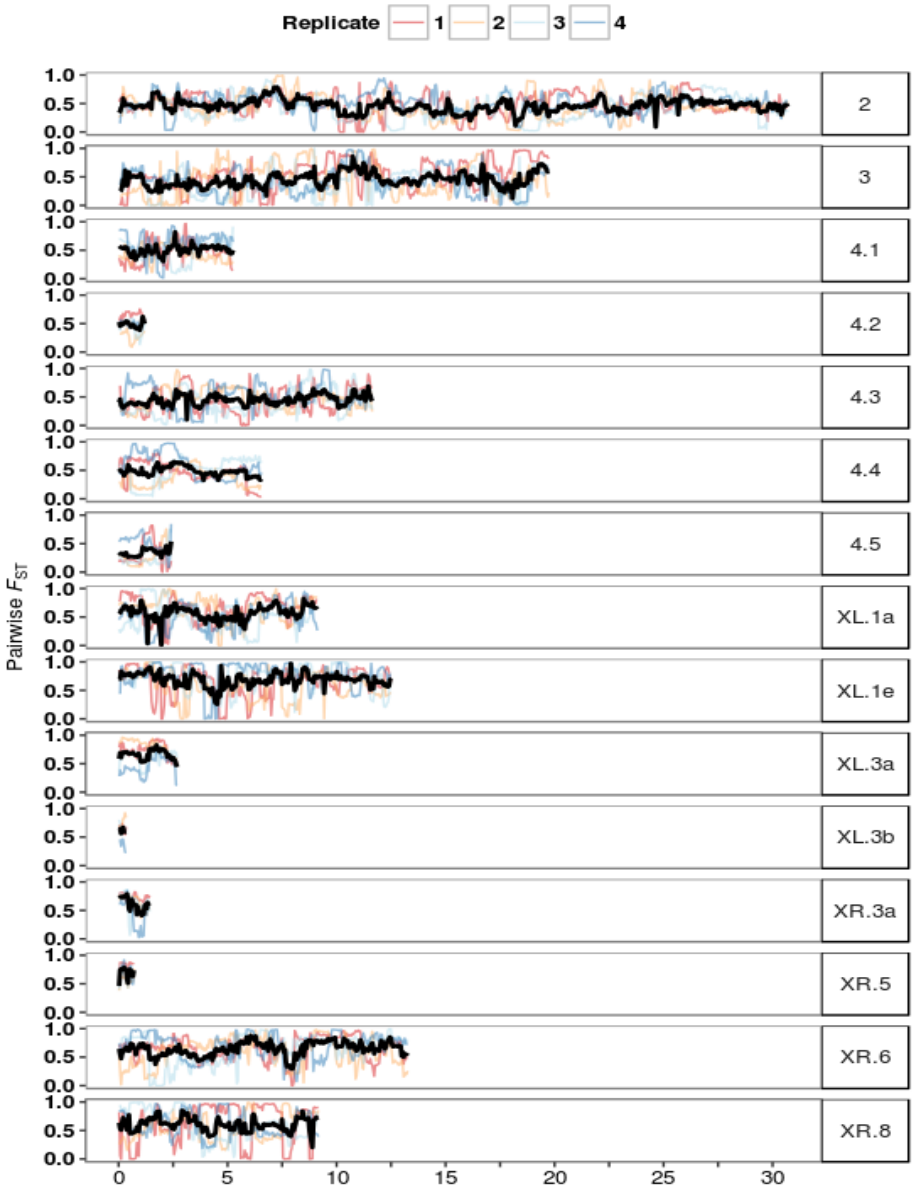


Figure 3.4. Pairwise F_{ST} calculated for each replicate pair of E and M experimental evolution treatment lines. F_{ST} is calculated for 50kb windows with a 5kb overlap between windows. Panels represent different chromosome regions. The X-axis represents position or distance along the chromosomal region. Coloured lines give the pairwise estimates for each pair of E and M lines. The black line shows the mean F_{ST} across all replicates. The *x*-axis gives the distance along each chromosome in Mb.

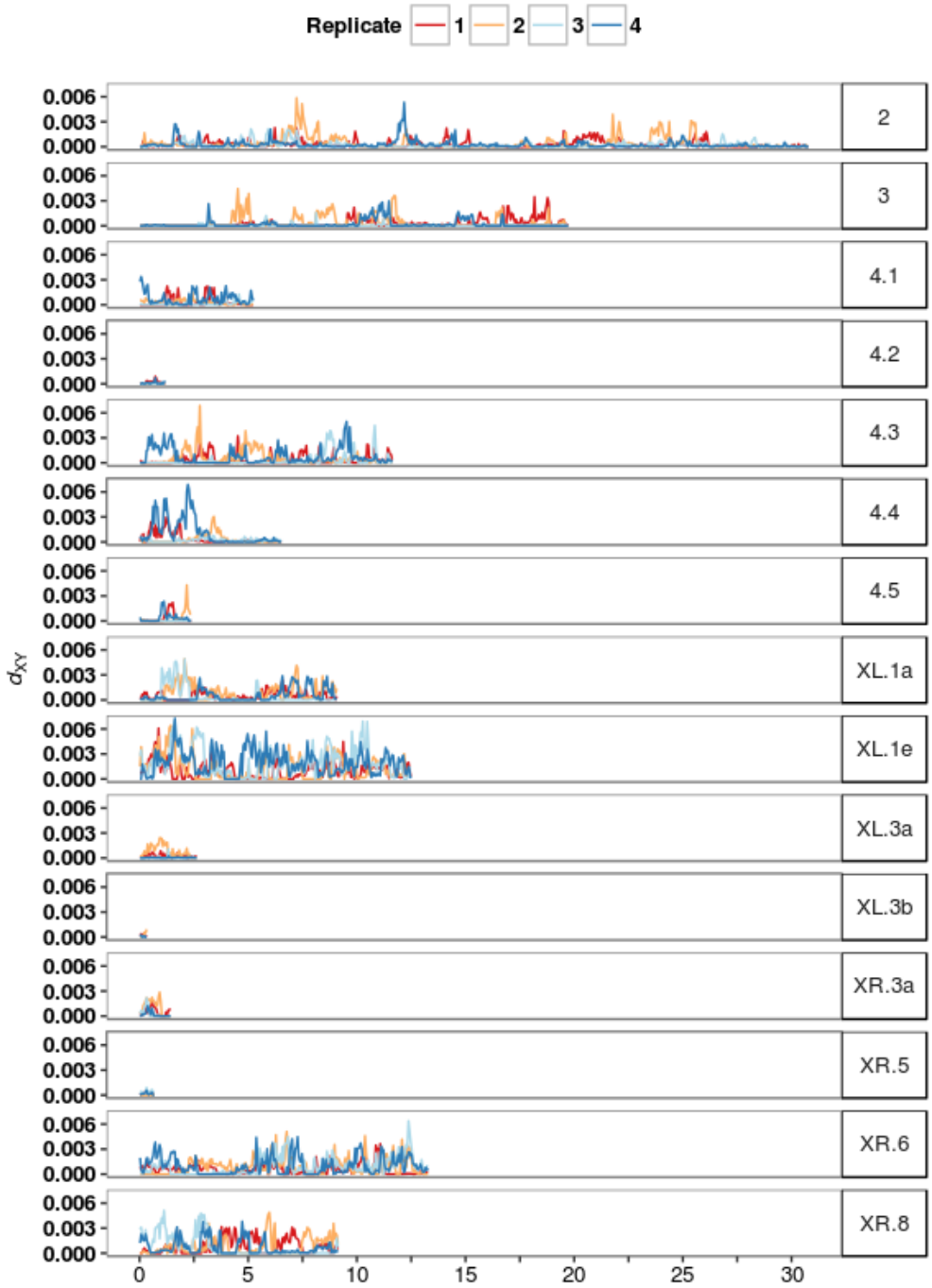


Figure 3.5. Pairwise d_{XY} between E and M lines for each replicate. The *x*-axis gives the distance along each chromosome in Mb

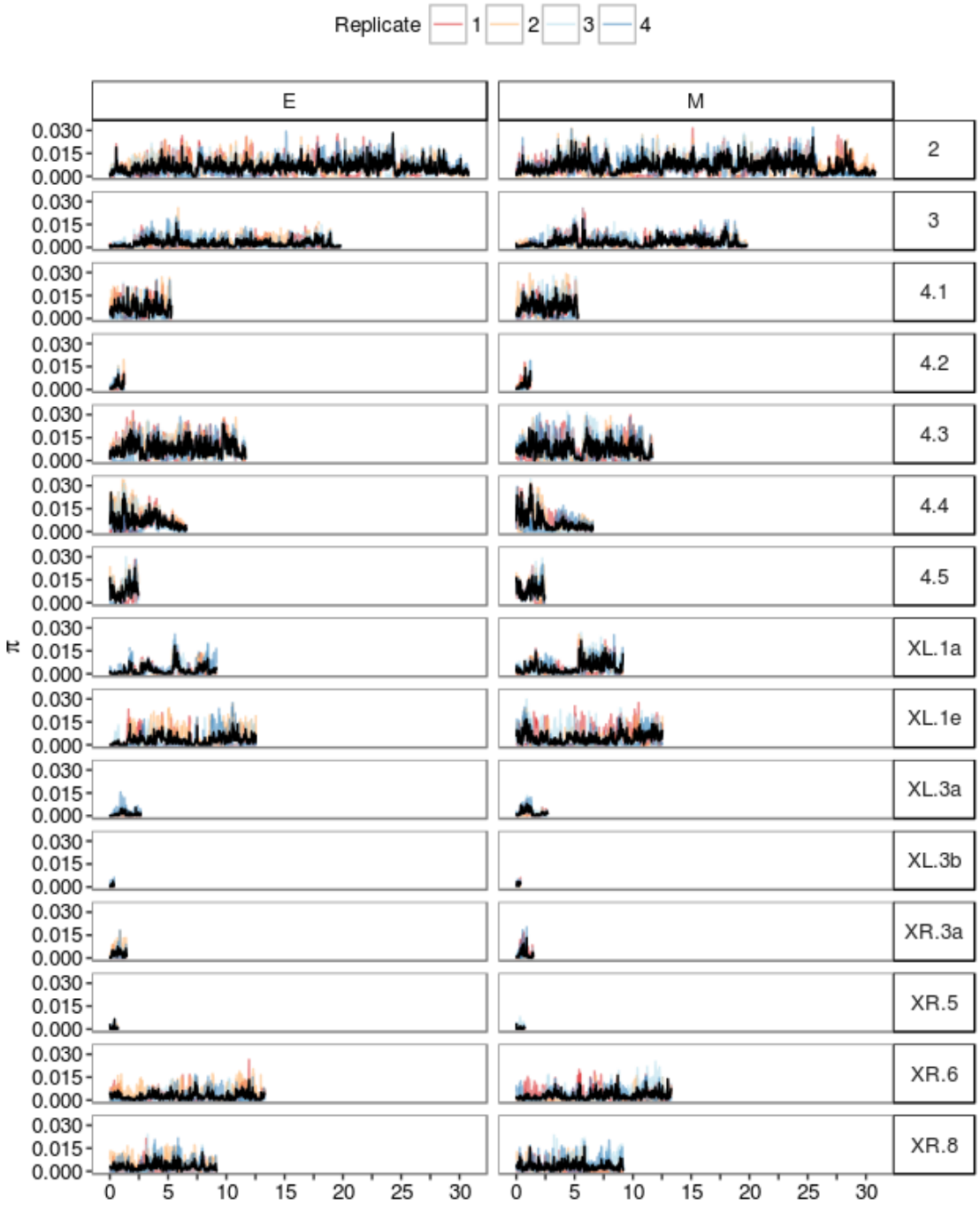


Figure 3.6. Estimates of nucleotide diversity (π) throughout the genome for each replicate in E and M lines. The black line gives the mean π across all replicates. The *x*-axis gives the distance along each chromosome in Mb

3.3.2 Identifying Candidate SNPs

In total, 1,130,944 SNPs were called and analysed. Many of these SNPs occur in intergenic regions with 652,101 of all SNPs (~63%) falling outside of annotated genes. The distribution of SNPs across the chromosomes is shown in figure 3.7. Using q-values to correct for multiple testing, 323 SNPs achieve genome-wide significance for a consistent difference in allele frequency between E and M treatments ($q < 0.05$, referred to as "candidate" SNPs). The genome-wide distribution of SNPs is not random even when accounting for the length of the chromosomes. There are more SNPs on the 2nd, 4th and X chromosomes than expected on the basis of their length, as a proportion of the total genome length, and fewer SNPs than expected on the 3rd chromosome (Chi-squared statistic = 60577, d.f. = 4, $p < 0.001$). By contrast, there are more candidate SNPs on the third chromosome and both arms of the X chromosomes than expected by on the basis of their lengths (Chi-squared statistic = 213.67, d.f. = 4, $p < 0.001$).

There are several clear "chimneys" or "peaks" of clustered candidate SNPs that show a consistent allele frequency difference between the E and M lines (figure 3.7). One obvious, dense peak of candidate SNPs stands out at one end of chromosome three (figure 3.7). Additional wide regions of candidate SNPs are seen in various regions of the X chromosome (figure 3.7), consistent with the higher differentiation seen on X chromosomes. It is notable that the patterns of genome-wide variation correspond to some degree to the results shown in figure 3.7. For example, average pairwise F_{ST} seems higher the region of condidate SNPs on chromosome three than elsewhere (figure 3.4). Similarly, two peaks of F_{ST} on the last group (group 8) of the X chromosome seem to align with the to regions covered by a few highly significant SNPs (figures 3.4 and 3.7). There are many known inversions on chromosome three in *D. pseudoobscura*. However, none of the known breakpoints seem to co-localise with the cluster of candidate SNPs (figure 3.8).

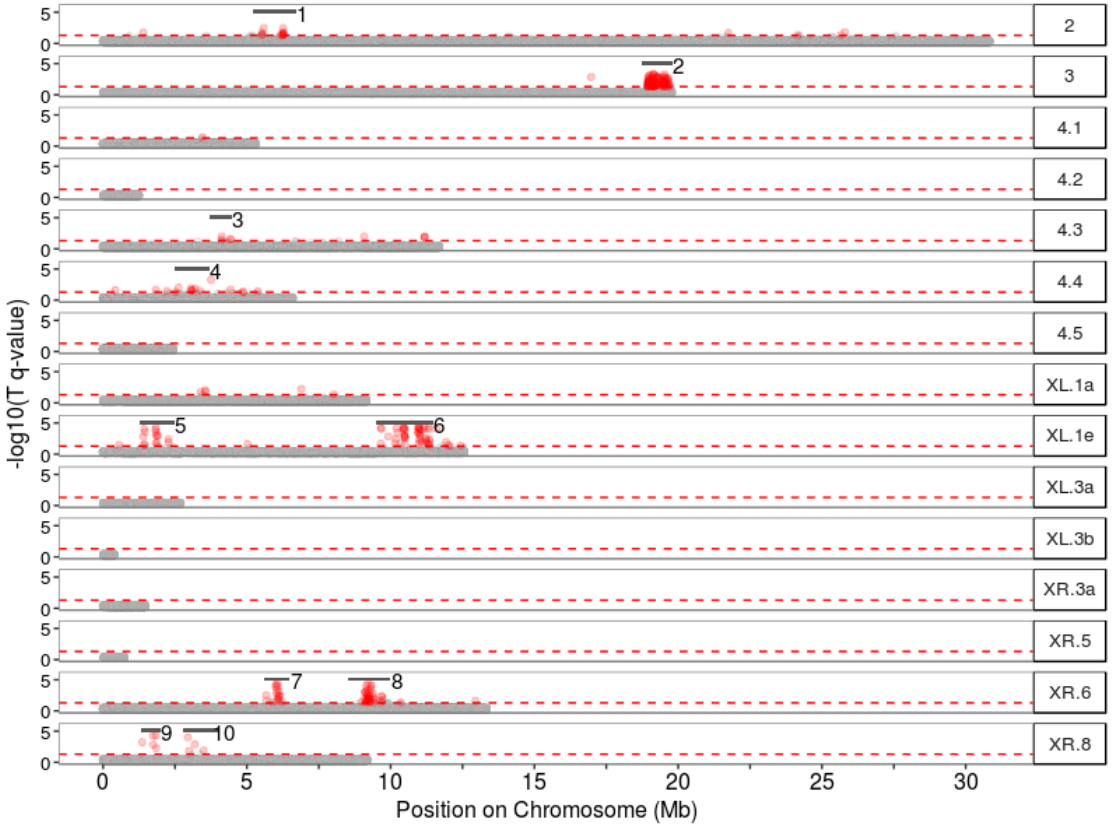


Figure 3.7. The distribution of analysed SNPs across the main chromosomes. The SNPs that have a q-value < 0.05 (see main text) are highlighted in red. Shown are the -log10(q-values) for the treatment (T) effect in a quasibinomial GLM. The black numbered lines delineate peak regions.

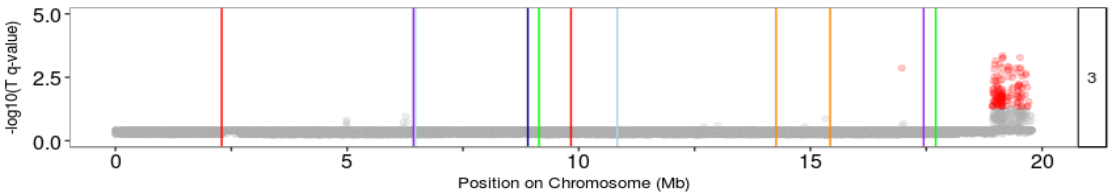


Figure 3.8. Positions of inversion breakpoints relative to the cluster of candidate SNPs on chromosome 3. Vertical lines on chromosome 3 show the break point positions of common arrangements (data and colours as in Wallace et al., 2011).

Figures 3.9 and 3.10 show the patterns of π and Tajima's *D* respectively within the regions with many candidate SNPs (highlighted in figure 3.7). Several of these regions also show reduced π and Tajima's *D* across all replicates in one or the other treatment (figures 3.9 and 3.10). In particular, the large region of top SNPs at the distal end of chromosome three shows reduced π in both E and M lines but strongly negative values of Tajima's *D* only in E lines suggestive of a sweep or stronger background selection within the E treatment. Other noteworthy regions are on the X chromosomes; XR.6 between 5.6 and 6.5 Mb, and XR.8 between 3.8 and 4 Mb (figure 3.10). Measures of genome-wide variation are sensitive to the resolution at which they are calculated and strong signals are probably required to observe an average pattern above the genomic background, "weaker" or more localised signatures are better identified by individual SNPs in the QB-GLM analysis.

3.3.4 Functional Analysis

In general, many candidate SNPs occur near genes (inlcuding within the coding region of a gene). There are 78 genes occurring within 1 kb of a candidate SNP. If the region is extended to 1 Mb the number rises to 138 genes. If the coding regions and any regions downstream of a gene is excluded, i.e. only SNPs upstream of a gene are counted, very few (~0.046% [15/167]) of the candidate SNPs occur within 1 kb upstream of a gene. By comparison, ~0.048% of all SNPs occur within 1 kb upstream of an annotated gene. If the upstream region is extended to 1 Mb the proportion of SNPs is higher (28%). In total, 142 (44%; q-values) SNPs occur within the coding region of 63 uniquely annotated genes.

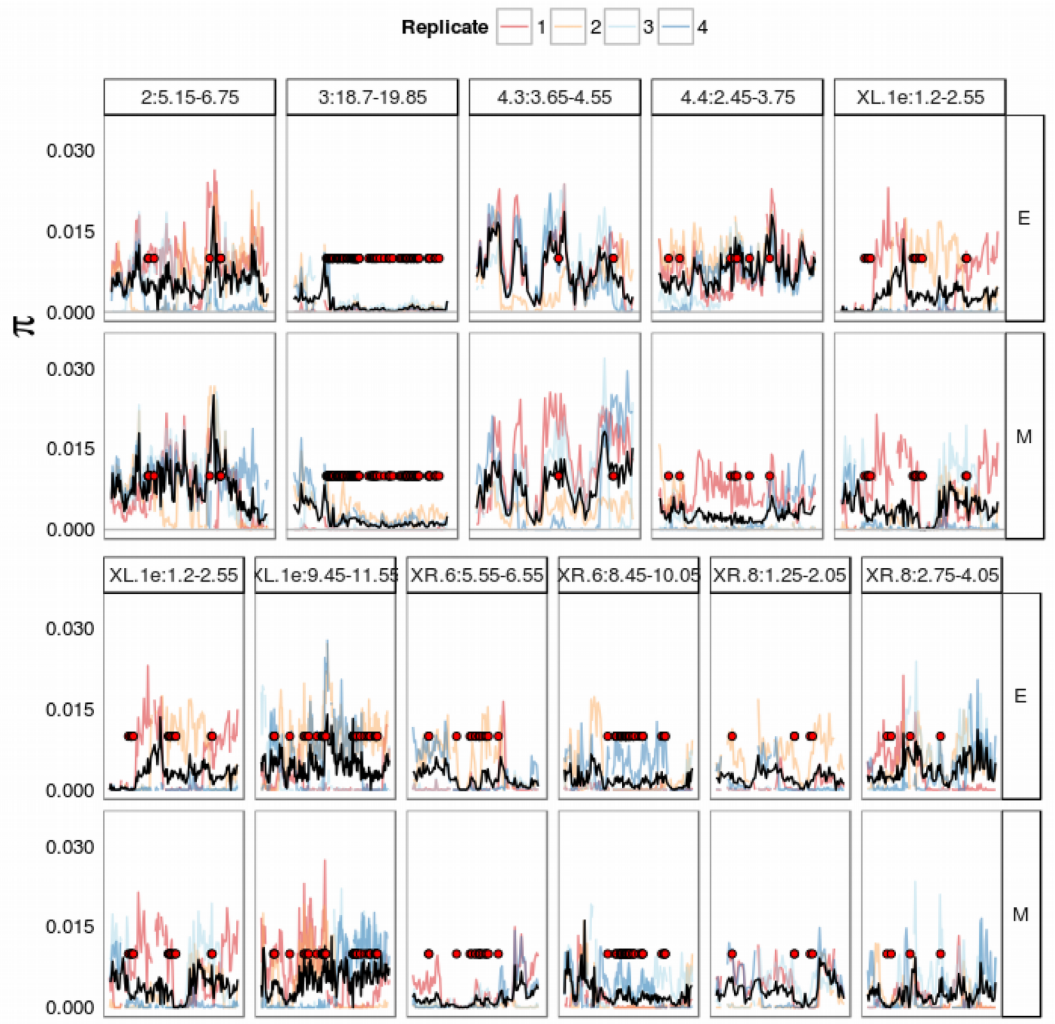


Figure 3.9. Estimates of π for windows within the regions identified in figure 3.7. Also shown are the top SNPs (q-value < 0.05) that occur within each region. Panel titles give the chromosomal locations of regions. Note that the *x*-axes are not to the same scale.

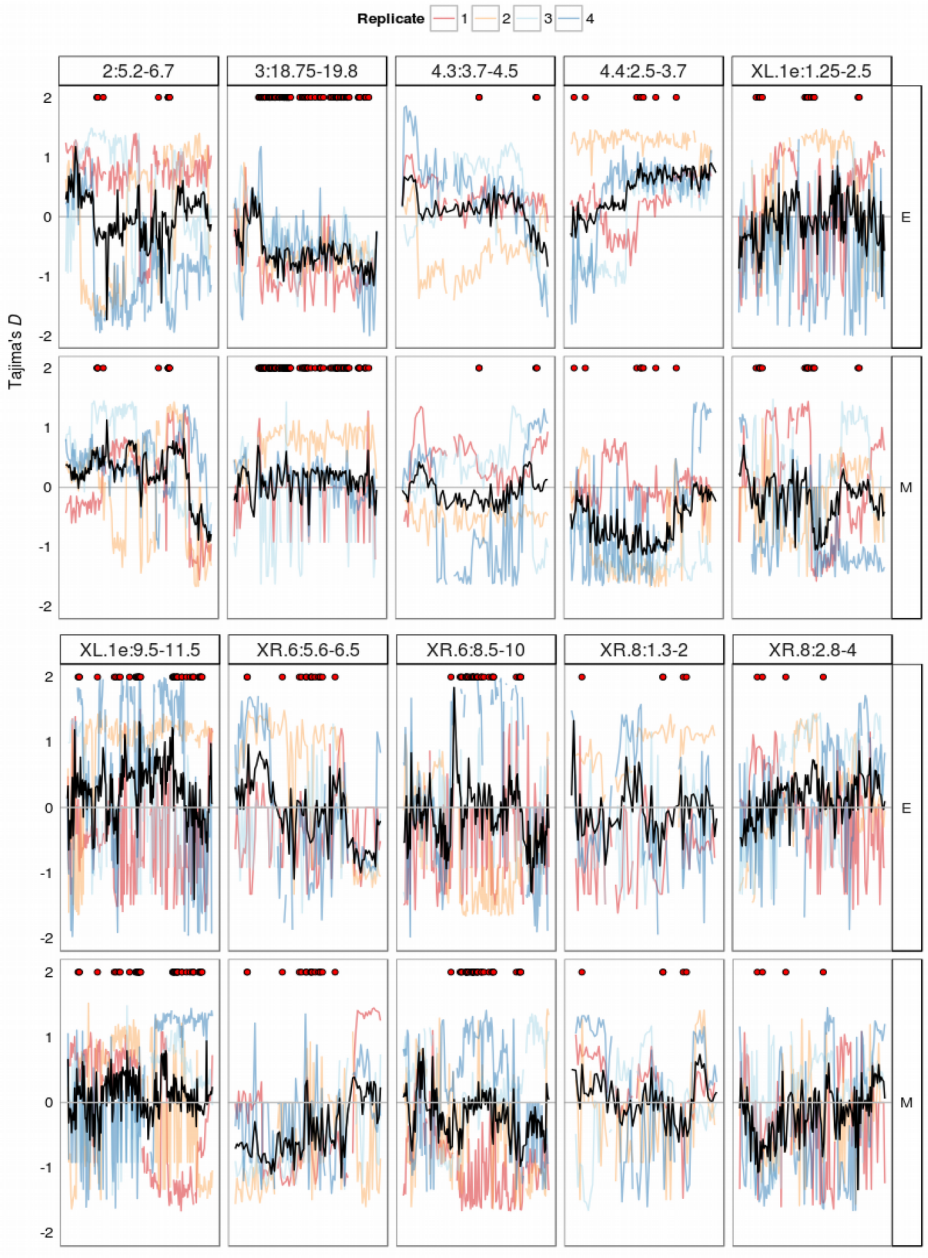


Figure 3.10. Estimates of Tajima's D for windows within the regions identified in figure 3.7. Also shown are the top SNPs (q-value < 0.05) that occur within each region. Panel titles give the chromosomal locations of regions. Note that the *x*-axes are not to the same scale.

Several genes generally occur within the wide peak regions identified above (see *3.3.3 Identifying Candidate SNPs*; Figure 3.7). In particular, the wide region on chromosome three contains 88 genes, 51 of which have some manner of annotation in the online database FlyBase (dos Santos et al., 2014) and 66 of which have a candidate

SNP within 1Mb (not including the coding region). Some genes (*Acp53C14c*, *Acp53Ea*, *Acp53C14b* and *Acp53C14a*) are annotated as accessory gland proteins which are components of the "cocktail" of seminal fluid proteins passed from males to females during copulation. Other genes include ion channel proteins (e.g. *ppk6*, *Pickpocket 6)* or odorant binding proteins (*Obp47a*). The other regions identified in figure 3.7 also contain many genes. Most of the genes in these additional regions have diverse annotated functions within FlyBase (see *3.4 Discussion*). Table 3.3 shows the set of genes which are associated with the candidate SNPs and includes all genes that have a candidate SNP within 1Mb of the gene (including within the coding region of a gene).

In a GO term enrichment analysis candidate SNPs (by bonferroni correction) are aassociated with 72 genes representing 1,091 GO sets, three of which were significantly enriched for member genes (adjusted p-value < 0.05; Table 3.4). The most significantly enriched GO terms refer to "potassium ion binding", "alkali metal ion binding", and "pyruvate kinase activity" (table 3.4). If the "genic" region in the above analysis is extended from 1kb to 1Mb up- or down-stream, no GO terms are significantly enriched.

Genes within 1Mb of the candidate SNPs (q-value < 0.05) are not enriched for genes within the phenotypic classes "courtship behavior defective" and "mating rhythm defective". Nor are they enriched for genes within the class "stress response defective". Though some phenotype terms achieve marginal significance in an enrichment analysis of all phenotypic classes available, none remain significant after Bonferroni correction for multiple testing.

Table 3.3. Genes which have a significant SNP (q-value < 0.05) within 1 Mb. Given are the chromosomes and the positions along the chromosomes and the *D. melanogaster* ID for each gene identified within the regions shown in figure 3.7. Genes highlighted in **bold** are discussed in the text. Genes denoted with a "*" are also found in the differential expression datasets (discussed in the text). Genes denoted with "+" have candidate SNPs within the coding region itself.

Chromosome	***Start***	***End***	***FlyBase ID***	***Name***	***Notes***
chrom2	1322185	1405385	FBgn0010113	hdc	+
chrom2	5526845	5529577	FBgn0278604	dmt	
chrom2	5530857	5533664	FBgn0037734	trbd	+*
chrom2	5586391	5591727	FBgn0038874	ETHR	+
chrom2	6138182	6139420	FBgn0039029	CG4704	
chrom2	6258602	6259899	FBgn0086253	rumi	+*
chrom2	6261369	6262388	FBgn0039020	CG17141	+*
chrom2	6268799	6270691	FBgn0039017	CG6985	+*
chrom2	21743937	21744764	FBgn0087005	rtp	
chrom2	24159758	24172033	FBgn0026620	tacc	+*
chrom2	25658685	25663510	FBgn0053512	dpr4	+
chrom2	**25788233**	**25797861**	**FBgn0017581**	**Lk6**	**+**
chrom3	16975321	16976566	FBgn0033653	CG13192	*
chrom3	**18936878**	**18937443**	**FBgn0033573**	**Obp47a**	
chrom3	18949890	18957102	FBgn0020236	ATPCL	+*
chrom3	18965505	18999327	FBgn0265045	Strn-Mlck	+*
chrom3	18999724	19003285	FBgn0034053	Cyp4aa1	+
chrom3	19004871	19005653	FBgn0034052	CG8299	+
chrom3	19006657	19011317	FBgn0033062	Ars2	+
chrom3	19012759	19014608	FBgn0033061	SmydA-5	+
chrom3	19014638	19015705	FBgn0033060	CG7849	
chrom3	**19020160**	**19022351**	**FBgn0034489**	**ppk6**	**+**
chrom3	19052469	19054843	FBgn0034157	resilin	
chrom3	19069409	19078799	FBgn0267002	unc-104	+*
chrom3	19088766	19089493	FBgn0034151	CG15617	
chrom3	19135041	19167763	FBgn0050463	CG30463	+

chrom3	19212763	19218778	FBgn0040505	Alk	*
chrom3	19224870	19228509	FBgn0034145	CG5065	
chrom3	19249061	19250239	FBgn0034144	CG5089	
chrom3	19254826	19259400	FBgn0034142	CG8306	+*
chrom3	**19268434**	**19274025**	**FBgn0086358**	**Tab2**	+
chrom3	**19279213**	**19281405**	**FBgn0034433**	**EndoB**	+
chrom3	19296634	19306333	FBgn0000008	a	+
chrom3	19335326	19336318	FBgn0029084	gom	
chrom3	19364579	19365585	FBgn0050281	CG30281	
chrom3	19367811	19373637	FBgn0243516	Vrp1	+
chrom3	19382049	19389507	FBgn0034693	CG11073	+*
chrom3	19445252	19447801	FBgn0022160	Gpo-1	+
chrom3	**19453912**	**19457254**	**FBgn0034046**	**tun**	
chrom3	**19475385**	**19479391**	**FBgn0003130**	**Poxn**	+
chrom3	19501239	19542746	FBgn0265991	Zasp52	+*
chrom3	19545467	19577664	FBgn0050089	CG30089	+*
chrom3	19630360	19647703	FBgn0023441	fus	+
chrom3	19734572	19735121	FBgn0034033	CG8204	*
chromXL.1e	1395149	1397023	FBgn0030828	CG5162	
chrom4.1	3483969	3484682	FBgn0028871	Cpr35B	
chrom4.3	4126443	4127132	FBgn0261523	CG42658	
chrom4.3	**4399026**	**4491686**	**FBgn0000636**	**Fas3**	+
chrom4.3	**9017139**	**9083795**	**FBgn0000497**	**ds**	+
chrom4.3	11189475	11190546	FBgn0031619	CG3355	
chrom4.4	388540	389775	FBgn0032457	CG15483	
chrom4.4	1810294	1811069	FBgn0031430	CG3528	
chrom4.4	2221448	2225787	FBgn0028370	kek3	
chrom4.4	2531613	2532249	FBgn0086691	UK114	
chrom4.4	2628466	2632798	FBgn0032856	CG16798	
chrom4.4	**3016820**	**3043507**	**FBgn0011676**	**Nos**	
chrom4.4	3079887	3088414	FBgn0031730	CG7236	+
chrom4.4	3177598	3206410	FBgn0051646	DIP-theta	+
chrom4.4	3370422	3371458	FBgn0021856	l(2)k14505	
chrom4.4	3765179	3767378	FBgn0264443	CG43861	

chrom4.4	4433137	4434869	FBgn0031462	CG2964	
chrom4.4	4870376	4900984	FBgn0028644	beat-Ic	
chrom4.4	**5279506**	**5353512**	**FBgn0261563**	**wb**	
chromXL.1a	3402094	3404275	FBgn0030417	CG15725	
chromXL.1a	3520958	3564973	FBgn0085446	CG34417	+
chromXL.1a	6885241	6902916	FBgn0028480	CG17841	+
chromXL.1a	8033492	8049410	FBgn0263511	Vsx1	*
chromXL.1e	549414	559856	FBgn0265767	zyd	+*
chromXL.1e	**11527037**	**11596083**	**FBgn0004198**	**ct**	+
chromXL.1e	11920786	11922143	FBgn0028665	VhaAC39-1	
chromXL.1e	**11991285**	**12006785**	**FBgn0259168**	**mnb**	+
chromXL.1e	12047353	12051483	FBgn0030869	Socs16D	
chromXL.1e	12464705	12470056	FBgn0263772	CG43689	
chromXL.1e	1427415	1430252	FBgn0030833	CG8915	*
chromXL.1e	1430865	1431545	FBgn0030834	CG8675	
chromXL.1e	**1448623**	**1468865**	**FBgn0262111**	**f**	+
chromXL.1e	**1815562**	**1819741**	**FBgn0000709**	**fliI**	+
chromXL.1e	1832624	1915270	FBgn0031174	CG1486	+*
chromXL.1e	2268633	2269605	FBgn0025644	CG14424	
chromXL.1e	2282530	2283510	FBgn0025645	CG3598	
chromXL.1e	**4932305**	**5100644**	**FBgn0052600**	**dpr8**	+
chromXL.1e	**9641088**	**9645320**	**FBgn0042650**	**disco-r**	
chromXL.1e	9919644	9921709	FBgn0010416	TH1	
chromXL.1e	10145547	10148164	FBgn0030369	Cyp318a1	
chromXL.1e	10172480	10174567	FBgn0030459	CG12723	+
chromXL.1e	10203322	10205107	FBgn0030456	CG4332	
chromXL.1e	10357747	10359420	FBgn0029854	CG3566	+
chromXL.1e	10440872	10445027	FBgn0030964	Pvf1	+*
chromXL.1e	10453739	10454782	FBgn0030963	CG7101	
chromXL.1e	10462034	10465902	FBgn0030974	CG7358	+*
chromXL.1e	10476195	10486631	FBgn0030976	CG7378	
chromXL.1e	10942690	11086962	FBgn0029939	CG9650	+
chromXL.1e	11142389	11145702	FBgn0029941	CG1677	
chromXL.1e	11183719	11193894	FBgn0029943	Atg5	+*

chromXL.1e	11221056	11236692	FBgn0029946	CG15034	+
chromXL.1e	11298567	11300021	FBgn0029946	CG15034	
chromXL.1e	11308498	11308815	FBgn0085366	CG34337	
chromXL.1e	11309771	11312536	FBgn0029950	CG9657	
chromXR.6	5676030	5682174	FBgn0036359	CG14105	+
chromXR.6	5903242	5904246	FBgn0011335	l(3)j2D3	*
chromXR.6	5989176	5997130	FBgn0263776	CG43693	
chromXR.6	6009943	6011469	FBgn0052081	CG32081	+
chromXR.6	**6058111**	**6060738**	**FBgn0013469**	**klu**	
chromXR.6	6069761	6070630	FBgn0036112	CG14147	*
chromXR.6	6153269	6153936	FBgn0036111	Aps	
chromXR.6	6189337	6205732	FBgn0026160	tna	
chromXR.6	**8994022**	**9008019**	**FBgn0052062**	**Rbfox1**	
chromXR.6	**9084845**	**9087358**	**FBgn0001179**	**hay**	**+***
chromXR.6	9094682	9109295	FBgn0052066	CG32066	+
chromXR.6	9145219	9148073	FBgn0036732	Oatp74D	
chromXR.6	9164395	9164724	FBgn0052185	edin	
chromXR.6	9179250	9179993	FBgn0036729	CG13733	
chromXR.6	9191311	9191750	FBgn0036726	QIL1	
chromXR.6	9206551	9210630	FBgn0036725	CG18265	
chromXR.6	9223288	9227185	FBgn0052176	CG32176	+*
chromXR.6	9230300	9231325	FBgn0052174	CG32174	+*
chromXR.6	9259642	9266610	FBgn0027660	blot	+
chromXR.6	**9303419**	**9304657**	**FBgn0015550**	**tap**	
chromXR.6	9331431	9333269	FBgn0036710	CG6479	
chromXR.6	9350219	9350956	FBgn0036706	ND-24L	
chromXR.6	9420888	9421520	FBgn0036704	CG6497	
chromXR.6	9432003	9437677	FBgn0035896	CG6983	
chromXR.6	9663038	9663932	FBgn0250815	Jon65Aiv	
chromXR.6	9675914	9682368	FBgn0052406	PVRAP	
chromXR.6	9730496	9739892	FBgn0035656	CG10479	*
chromXR.6	10357879	10360022	FBgn0036155	CG6163	
chromXR.6	12926872	12929590	FBgn0036702	CG6512	
chromXR.8	1354776	1355942	FBgn0035124	ttm2	

chromXR.8	1737338	1743845	FBgn0035798	frac	+
chromXR.8	1825052	1826996	FBgn0036576	CG5151	
chromXR.8	2945576	2947737	FBgn0035575	CG7509	
chromXR.8	2996331	3006274	FBgn0052237	CG32237	
chromXR.8	3168594	3169206	FBgn0035585	ATPsynCF6L	
chromXR.8	3489755	3491254	FBgn0035643	CG13287	+

Table 3.4. Gene Ontology (GO) categories which had a significant over-representation among genes within 1 kb of SNPs with a q-value < 0.05. Analysis was performed in Gowinda (Kofler & Schlötterer, 2012). Comparing the candidate SNPs to the background of all discovered SNPs. All FDR corrected p-values < 0.05. GO categories are given in decreasing order by level of significance.

GO term ID	***P-value (adjusted)***	***GO term Description***	***Genes Found among top SNPs***
GO:0030955	0.00004 (0.021)	potassium ion binding	FBgn0031462, FBgn0036723
GO:0031420	0.00004 (0.021)	alkali metal ion binding	FBgn0031462, FBgn0036723
GO:0004743	0.00004 (0.021)	pyruvate kinase activity	FBgn0031462, FBgn0036723

Transcription factor (TF) binding site motifs for *fru* and *dsx* are not significantly enriched around the candidate SNPs in comparison to the genomic background. This also holds if only regions around top SNPs that lie within 1 Mb upstream of a gene are considered and if the control set of sequences is left out. None of the other TF binding motifs within the MEME database of motifs are significantly enriched after correction for multiple testing.

Of all the genes within 1Mb of the top SNPs (q-value < 0.05) 32/138 (23%) occur within the dataset of differentially expressed (DE) genes identified by Immonen et

al., (2014). This enrichment is greater than expected by chance (empirical p-value = 0.006, 10,000 bootstraps). Overlap with the Veltsos et al., (*in prep.*) dataset is lower with only 5 (~0.04%) of the genes near top SNPs occurring in the DE set, which is not greater than expected by chance (empirical p-value = 0.47, 10,000 bootstraps). Only two genes occur in all three datasets both of which are annotated as uncategorised proteins.

3.4 Discussion

A major goal in evolutionary biology is understanding the process of sexual selection and its interplay with sexual conflict in different mating systems. Recently adopting a genomic approach has started to provide more detailed understanding of these processes both in terms of the physical loci and their organisation throughout the genome, as well as their effects on the phenotypes through gene expression (Ritchie & Butlin 2014, Wilkinson et al., 2015). Experimental evolution studies have historically been a useful tool to study sexual selection and conflict under different mating systems because they naturally lend themselves to manipulation of the social environment and, hence, the mating system. I conducted a genome-wide pool-seq study of an ongoing long-term experimental evolution study in *D. pseudoobscura* which was set up in 2002 to study the interplay of sexual selection and sexual conflict and which has produced many intriguing results, many of which match predictions from theory.

A survey of genome-wide variation and differentiation between replicated treatment lines shows that F_{ST} is high and quite variable (figure 3.2) possibly due to drift and/or small population sizes. Similarly, nucleotide diversity is quite variable though with the notable region on chromosome 3 as an extended region of consistently low diversity (figures 3.4 and 3.9). Meanwhile, patterns of F_{ST} and diversity on the autosomes and X chromosomes show differences between E and M lines. Polyandry is expected to further reduce the effective population size (N_e) on X chromosomes relative to autosomes which should in turn drive down diversity. Indeed diversity is on average lower across E treatment lines but the difference is not significant. However, the difference in F_{ST} between the X and autosomes is significantly greater in E lines than in M lines. This difference is greater than expected from differences in N_e alone suggesting

an effect of selection. This is also consistent with the disproportionate number of candidate SNPs and reduced diversity on the X chromosomes. These results suggest a faster-X and a greater efficiency of selection on the X chromosome due their hemizygosity in males (Vicoso & Charlesworth 2006; Vicoso & Charlesworth 2009). However, sexual selection and conflict is expected to increase the X:A ratio of diversity (Mank et al., 2014; Vicoso & Charlesworth 2009), suggesting that these forces are not driving the pattern seen here. In sum, the results suggest that polyandry, and the mating system in general, can have a great influence on the evolution of sex chromosomes (Vicoso & Charlesworth 2006; Ellegren 2009; Corl & Ellegren 2012). X:A ratios of diversity and F_{ST} have been reported in XY systems (Ellegren 2009; Arbiza et al., 2014), but explicit comparisons of the ratios across different mating systems have only been carried out in avian ZW systems (Borge et al., 2005; Corl & Ellegren 2009). These results thus provide valuable insight into the generality of the effects of mating systems on sex chromosome evolution.

Given that the differences on the X and autosomes are generated by the dynamics of differences in N_e it would be of interest to estimate N_e directly from the molecular data. This was done from microsatellite loci at earlier generations (Snook et al., 2009) but this study had only four loci and none were available on the X-chromosome. Although many methods for estimating N_e from molecular data exist (e.g. Wang 2005) the details of pool-seq design experiments introduce some difficulties (Jónás et al., 2016). Some methods have been developed for pool-seq that use the change in allele frequency during experimental evolution and the variance (e.g. Foll et al., 2014; Reed et al., 2014; Jónás et al., 2016). But these methods would require more sequencing of earlier, frozen samples from the experimental evolution lines. Alternatively, N_e can be estimated from current genetic diversity at neutral markers (Wang 2005). This relies on identifying neutral markers which should be possible, to some extent, in the datasets presented here. Alternatively, the lack of neutrality can to some extent be controlled for by estimating N_e first in windows, then averaging these windows across the genome, followed by a bootstrapping approach to obtain confidence intervals.

This study also identifies sets of SNPs that show consistent allele frequency differences between the E and M experimental evolution treatment lines across

biological replicates. The locations of these loci correspond fairly well to regions of high differentiation (F_{ST}; Figure 3.2) and higher divergence (figure 3.3). Meanwhile, close inspection of the regions of candidate SNPs show that many of them coincide with reductions in π (figure 3.9) and Tajima's *D* (figure 3.10) which is an indication of selective sweeps (Huber & Lohmueller 2016). Taken together, these observations offer compelling evidence for selection as the driver of allele frequency changes in these regions. Interestingly, some of the regions show stronger evidence of selection within the M lines than the E lines. For example, regions on the 4th (chromosome 4.4: 2.5-3.7 Mb) and X chromosomes (XL.1e: 1.25-2.5, 5.6-6.5, and 2.8-4 Mb) show reduced Tajima's *D* in M lines. This raises the question of which treatment imposes more of a divergent selection pressure on the population. *D. pseudoobscura* is a naturally polyandrous species and while the elevated rates of mating in E lines are designed to be higher than in natural populations it may be that the imposition of a monogamous mating system is a greater change. These results may also explain why the ratio of F_{ST} on the X and autosomes is higher than expected from differences in N_e even in M lines.

The large cluster of candidate SNPs on chromosome three show a characteristic signal expected from a large haplotype block of many linked SNPs such as an inversion. Several known inversions and genomic arrangements are known on chromosome three in *D. pseudoobscura* (Sturtevant & Dobzhansky 1936; Dobzhansky & Sturtevant 1937; Wallace et al., 2011). The breakpoints of the most common arrangements have been mapped (Wallace et al., 2011) and are shown in figure 3.6. However, the observation that none of these breakpoints align exactly with the observed cluster, coupled with the fact that the original stock for the experimental evolution lines was from Tucson, Arizona where the "Arrowhead" arrangment is the most common (Patton et al., 1966), perhaps suggests that this region is unlikely to be one of the well known inversions. However, other, low frequency inversions are known to segregate within *D. pseudoobscura* and a more thorough investigation of these lines is required to rule these out.

Many of the genes that occur within the regions delineated by the peaks of candidate SNPs in figure 3.6 have mutation and phenotype information (from FlyBase) that indicate roles in courtship, mating, fertility and feeding or circadian rythms. For example, four of the genes under the peak on chromosome three (figure 3.2; table 3.3;

Acp53C14c, *Acp53Ea*, *Acp53C14b* and *Acp53C14a*) have been experimentally characterised as mating plug proteins in *D. melanogaster* (Avila et al., 2015). Furthermore, the mates of *D. melanogaster* males in which either *Acp53C14a* or *Acp53C14b* are knocked down show reduced fertility in terms of the number of eggs laid than the mates of wild-type males (Avila et al., 2015). Variants of the gene *nonA* (no-on-transient A) encode species specific variation in various song behaviours (Campesan et al., 2001), and *nonA D. melanogaster* mutants show abnormal courtship song (Rendahl et al., 1992). Meanwhile, *Cyclic* (*cyc*) and *Clock* (*clk*) are important regulators of the circadian rhythm (Williams & Sehgal 2001; Sokolowski et al., 2001). A more robust set of genes lying within these regions are those that also have a candidate SNP (q-values < 0.05) within 1 Mb (table 3.3). Several of these genes have roles in the expression of phenotypes (e.g. male manipulation of female) that are directly or potentially related to phenotypic responses to selection seen in previous studies of these experimental evolution lines.

The identified genes which were related to courtship include *Odorant-binding protein 47a* (*Obp47a*), which is part of a family of similar proteins that are important in the detection of chemical stimuli in courtship and foraging (Hekmat-Scafe et al., 2002), this family of genes is also known to be under rapid molecular evolution within the *Drosophila* clade (Hekmat-Scafe et al., 2002; Vieira et al., 2007; Gardiner et al., 2008; Vieira & Rozas 2011) and dynamic gene family evolution across arthropods in general (Vieira & Rozas 2011). Another gene within the wide region of candidate SNPs on chromosome three, *ppk6*, is a member of the sodium ion channel family of *ppk* proteins (Zelle et al., 2013; Ben-Shahar 2011). Other members of this family, e.g. *Ppk23,* are expressed in sexually-dimorphic sensory bristles on the front legs. Mutations or disruptions in the expression of *ppk23* delays and reduces the amount of male courtship (Lu et al., 2012; Thistle et al., 2012; Toda et al., 2012). This gene is also involved in the detection of female pheromones (Lu et al., 2012; Thistle et al., 2012; Toda et al., 2012).

Another gene, *disconnected*-related (*disco-r*), is similar in sequence and in some known functions to the gene *disco. disco* is thought, from mutant phenotypes, to be involved in antennal (Day et al., 2009) and brain (Blanchardon et al., 2001) development as well as eclosion rythms (Williams & Sehgal 2001; Sokolowski et al., 2001). *disco-r* itself has also been associated with antennal development (Patel et al.,

2007). The gene *tab2* (*TAK1-associated binding protein 2*) was one of several genes found near an enhancer trap insertion that caused courtship song abnormalities in a study of *Drosophila melanogaster* (Moran & Kyriacou 2009).

Yet another example of a courtship related gene is *poxN* (*Pox neuro*), a transcription factor of the *pox* gene family, for which mutants show abnormal courtship bahavior which is likely due to aberrant development of wing and leg chemosensory bristles, as well as changes to genital structures (Boll & Noll 2002). It is noteworthy that *ppk23* expression is reduced in the appendages of *poxN* mutants lacking which lack normal chemosensory appendages (Lu et al., 2012). Meanwhile, knockdown of *Target of poxN* (*tap*) increases glucose levels in the heamolymph, suggesting that *poxN* may be important in regulating resource allocation (Ugrankar et al., 2015).

A particularly intriguing set of genes identified are related generally to the production and preferences of courtship song, which is produced by wing movements in *Drosophila* species. Variants of the gene *Forked* (*f*) show reduced "sound-evoked" electrical potentials in the antennae suggesting a reduced auditory response and defective hearing (Cosetti et al., 2008). *Klumpfuss* (*klu*) is a transcription factor implicated in the regulation of larval feeding bahavior (Melcher & Pankratz 2005) but also the development of auditory sensory organs (Hu & Castelli-Gair 1999; Kaspar et al., 2008). Meanwhile, *Flightless I* (*fliI*) mutants show a reduction in flight ability (Homyk & Sheppard 1977). More recent work with RNAi suggests that this reduction is due to mutations causing irregular development of "indirect flight muscles" that control wing movements (Schnorrer et al., 2010). Nitric oxide synthase (*Nos*) is important in the development of organs from imaginal discs. Inhibition of *Nos* in larvae results in enlarged adult structures such as wings and leg segments (Enikolopov et al., 1999). Furthermore p-element insertions near the gene *wing blister* (*wb*) results in significant differences in wing shape with respect to wild-type flies indicating a role in the development of wing shape (Carreira et al., 2011). The gene *dachsous* (*ds*) is required for pattern formation in the imaginal wing disc (Rodríguez 2004) and influences adult wing shape (Baena-Lòpez 2005). Mutations in *minibrain* (*mnb*) reduced brain volume (Sokolowski 2001) but also functions with *ds* and *riq* (*riquiqui*) to control growth in the imaginal discs of developing pupae (Degoutin et al., 2013). The gene *Cut* (*ct*) is a transcription factor which is expressed in the developing spiracle chamber (Hu &

Castelli-Gair 1999). *cut* mutants show abnormal development of auditory organs and exhibit a lower sound-evoked neuronal potential than wild-type flies (Ebacher et al., 2007). *cut* is also required for the proper development of wing margins and several other phenotypes (Thumm & Kadowaki 2001).

Size, aggression, and mating outcomes are closely related. E males are on average slightly larger and court more than M males (Debelle et al., 2016), and development times are reduced among *D. melanogaster* males evolving under increased sexual selection (Hollis et al., 2016b). These phenotypes, too, are represented among the functions of genes near candidate SNPs. *haywire* (*hay*) is a regulator of growth and cell proliferation (Lee et al., 2015). *Lk6 kinase* (*Lk6*) shows increased expression in cyc^{01} mutants that are sleep deprived or starved (Thimgan et al., 2015), and is also important in the growth and development of *Drosophila* and influences adult body size, some *lk6* mutants are smaller than wilde-type (Arquier et al., 2005). Related to feeding behavior and growth, the gene *defective proboscis extension response 8* (*dpr8*) is a protein with very high similarity to *dpr1* which is required for the wild-type aversion to salt, the phenotype by which this gene is named (Nakamura et al., 2002). Meanwhile *dpr6*, another gene from this subfamily, is implicated in male agression, an important component of the ability of a male to gain access to mates either by coercion of females or competition among males (Shorter et al., 2015). Other genes are involved in female fecundity, another phenotype that is known to have shown a response to selection in these treatment lines. *Endophllin B* (*EndoB*) is required for proper egg-yolk upatake by oocytes (Tsai et al., 2014). Finally, several genes involved in memory and learning are represented near the top SNPs. *Fasciclin 3* (*fas3*) mutants show reduced memory (Dubnau et al., 2005). Overexpression of RNA-binding Fox protein 1 (*Rbfox1*; also known as Ataxin-2 Binding Protein 1; *A2BP1*) or suppression of an inhibitor microRNA (*miR-980*) enhances memory in *Drosophila* (Guven-Ozkan et al., 2016). It is also involved in the normal development of ovarian cysts (Tastan et al., 2010), and in the formation of wings (Usha & Shashidhara 2010). Finally, *Tungus* (*tun*) is involved in olfactory learning and memory (Dubnau et al., 2005)

One caveat to the above results is that the association of these genes with the candidate SNPs will be affected by assumptions about the locations of enhancers and promoters with respect to the genes they regulate. Here, I have used a region spanning

1Mb which I justify above. This is a relatively wide window and in many cases several genes occurin within 1Mb of a focal SNP, I have focused my analysis of the closest gene on the assumption that these are the most likely targets of regulation.

Nevertheless, many of these genes are known to influence traits that have shown a response to selection in these experimental evolution lines, or they influence traits that could provide a mechanistic basis for these changes (e.g. wing shape or wing muscle and courtship song). In addition, some genes have been implicated in more than one of the phenotypes of interest (e.g. *cut*) raising the possibility that observed phenotypic changes in the experimental evolution lines could be due to pleiotropic effects of changes within single genes. This could explain, for example, how both preference and song characters could change within the M treatment where no sexual selection (through mate choice) is present and so females should pay a cost to being choosy (Debelle et al., 2014). Relaxed selection on song characters alters average song characters within the treatment and pleiotropic effects on the development of auditory organs might result in female "preference" or bias for different courtship song characters. Such pleiotropic effects underlying the correlation of preferences and display traits are implicated from the song and preferences of Hawai'ian crickets *Laupala paranigra* and *L. kohalensis* (Shaw & Lesnick 2009; Wiley et al., 2011), wing colour and preference in *Heliconius* butterflies (Kronforst et al., 2006; Merrill et al., 2011), and in pheromone production and recognition in *Drosophila melanogaster (*Marcillac et al., 2005; Bousquet et al., 2012).

Previous work has shown that expression levels for sex biased genes change in response to experimental evolution under altered mating systems (Immonen et al., 2014; Hollis et al., 2012). This pattern may be driven by changes in enhancer or promoter regions that act in *cis* to control the expression of target genes. If this is true then markers which show consistent differences in allele frequencies between the two treatment should occur near genes that also show DE patterns. The comparison of the genes near the top SNPs in this study with two sets of DE genes identified from virgin female whole bodies in one study (Immonen et al., 2014) and from various different tissues, mating status and sexes in another study (Veltsos et al., *in prep*) suggest that there is some overlap between the sets. Many of the genes discussed above as occurring near candidate SNPs are also known to show some differential expression between

experimental evolution treatments lines. For example, *hay (haywire)*, which is a regulator of growth (Lee et al., 2015) is differentially expressed between virgin E and M females (Immonen et al., 2014).

However, transcription factor (TF) binding motif enrichment analysis does not find evidence for significant enrichment of TF motifs around candidate SNPs. This could be due to a lack of power with so few candidate SNPs and TF motifs relative to the background set. Alternatively, variants could lie in as-yet unknown binding motifs. Another possibility is that other sources of regulatory variation could be more important. Non-coding RNAs (Rinn & Chang 2012; Guttman & Rinn 2012) and microRNAs (miRNAs; Mohammed et al., 2017) have recently become recognised as important contributors to gene regulation at various stages (pre-/post-transcription and translation). However, characterisation of their function is still difficult and datasets of putative non-coding RNAs are only starting to become available in non-canonical model systems like *D. pseudoobscura* (Nyberg & Machado 2016; Mohammed et al., 2017). Considering these loci will become important in attempts to understanding the regulation of gene expression. Many lncRNAs, for example, seem to show sex-biased expression patterns (Nyberg & Machado 2016). These loci are also known to influence mating and courtship behaviour. In *D. melanogaster* the microRNA *miR-124* is important in proper production of pheromones and females prefer males with the wild-type *miR-124* locus over mutants (Weng et al., 2013)

This study has identified some potential targets for future knockdown validation using, for example, CRISPR-*cas9*. In particular, the observations that courtship song, which is produced by rapid wing movements, has diverged between these lines (Snook et al., 2005), as well as the finding of several genes that seem to be involved in wing muscle development and flight ability (*fliI, ds*, *cut*), hint at an important courtship trait for which the genetic basis may be amenable to investigation. Similarly, genes that code accessory gland proteins are found within the wider genomic peak on chromosome 3. Such proteins are prime candidates for a genetic basis of changes in female fecundity or costs of multiple mating. Additionally, several of the associated genes are known to be involved in various phenotypes that have not previously been scored or assayed in these experimental evolution lines (e.g. wing shape or differences in eye morphology). It would be of interest to quantify whether any difference in eye morphology is apparent

between the E and M treatments or if any differences in flight ability, wing shape, or wing muscle strength that are likely to be important correlates of courtship song. Similarly, since gene expression profiles can be drastically different across sexes, tissues and developmental stages, studies that analyse whole bodies or very large body regions may miss many patterns of differential gene expression. It would therefore be of benefit to determine if the genes identified here are DE in other body regions or tissues (e.g. developing or adult wing muscles, imaginal wing-discs) and could therefore account for differences in phenotypic traits (e.g. courtship song) seen across these treatment lines.

These results bear on the debate of what kinds of genome-wide patterns are expected as populations experience divergent selection. The experimental evolution lines do not have any gene flow between diverging populations. The landscape of *Ne*, and recombination, throughout the genome should be the same in each treatment line. Thus the only driver of differences in various population genomic statistics should be selection or drift. The generally high levels of F_{ST} seen in this study may represent the relatively low effective population size within treatment lines and an effect of neutral drift throughout the genome. However, F_{ST} does not capture the effect of selection very well; regions containing SNPs with consistently different allele frequencies do not generally show obviously higher F_{ST} (peaks).

The use of F_{ST} peaks alone as indicators of adaptively important regions in diverging populations (so-called "islands of speciation" harbouring "speciation genes" which contribute to reproductive isolation) has come under criticism (Noor & Bennett 2009; Cruickshank & Hahn 2014; Wolf & Ellegren 2017) and appears to be a very coarse measure of differentiation. F_{ST} has become a popular statistic used to infer regions showing barriers to gene flow in diverging populations with ongoing hybridisation and gene flow (Noor & Bennet 2009; Wolf & Ellegren 2017). However, peaks of F_{ST} ("islands of speciation") can be produced simply by restricted recombination, such as those in inverted regions (either neutral or selected; Noor & Bennet 2009). Reduced recombination generally reduces nucleotide diversity within a species or population and can inflate relative measures of differentiation (e.g. F_{ST}). Indeed, reanalysis of several recent studies that inferred such "islands" on the basis of localised elevated F_{ST} suggest that these patterns are better explained by reduced

diversity in these regions (possibly a result of selection and hitchhiking; Cruickshank & Hahn 2014). A solution may be to use more explicit coalescent modelling of different scenarios of ancestral variation, gene flow and selection (Noor & Bennet 2009). Alternatively, other population genetic statistics, such as Tajima's *D* or Fay and Wu's *H* which can be estimated from population genomic data can provide more information about selection acting within populations, however, these are rarely used to complement estimates of F_{ST} (Wolf & Ellegren 2017)

In the results above I contrast the measures of F_{ST} (figure 3.4) with estimates of d_{XY} (figure 3.5), π (figure 3.6 and 3.9), and Tajima's *D* (figure 3.10). One hypothesis that may be directly adressed by the above results is; if "islands" of speciation are generally not due to selection with ongoing gene flow, but instead due to selection at some loci coupled with non-assorted ancestral variation in the rest of the genome, then we should also see such peaks of F_{ST} in these experimental evolution studies. In this study the combination of π and Tajima's *D* as indicators of selective sweeps or background selection appears to be a fruitful approach in this case showing close concordance with individual SNP based tests of consistent allele frequency changes. Meanwhile, d_{XY} and F_{ST} do not obviously delineate these regions of SNPs, though the regions with clusters of candidate SNPs do show some hints of elevated F_{ST}. The lack of obvious peaks in F_{ST} could be due to a combination of the generally high F_{ST} throughout the genome and the relatively short timescale of the experiment (relative to evolution in natural populations). Clearly, peaks of "differentiation" can arise in the absence of gene flow. These regions of highly clustered candidate SNPs and slightly elevated F_{ST} may be regions of generally low recombination or even of inversions, but more data will be require to test this.

3.5 Concluding Remarks

The interaction between sexual selection and conflict in driving biological diversity is an important topic of study in evolutionary biology. A large part of mating systems is the female re-mating rate. Differences in the female re-mating rates produce differences in the levels of sexual selection and conflict. A fruitful approach to the study of mating systems has been to control access of females to males in experimental

evolution studies. Here I analyse pooled whole-genome sequence data from an ongoing experimental evolution study. The aim was to identify SNPs that show consistently different allele frequencies between experimental evolution treatments. I also used population genetic summary statistics to examine if these regions show evidence of selection. The results show several regions indicative of selective sweeps or background selection that co-localise with individual SNPs with consistent allele frequencies between treatment lines. Several genes near the candidate SNPs have phenotypic information established from mutant assays. Several of these phenotypes are directly or indirectly related to responses to selection in these same experimental evolution lines. In sum, the study has identified several promising loci that underlie phenotypic divergence in these experimental evolution lines. Further functional validation will require knockdown or knockout experiments.

Chapter 4 Identifying genomic markers associated with the female re-mating rate in *Drosophila pseudoobscura.*

Abstract

Identifying genomic loci associated with a trait of interest is the first step in many genetic studies. Many methods of genotype-phenotype association studies require large sample sizes or breeding designs that are often prohibitively difficult, especially in non-model organisms. As sequencing costs have fallen novel applications of sequencing methods are being developed to find efficient ways of identifying these associations. Such innovations have led to great advances in the analysis of, for example, experimental evolution studies and population genomics.

This chapter applies a novel approach in identifying SNP loci associated with the female re-mating rate in *Drosophila pseudoobscura*. The re-mating rate, a measure of how willing a female is to mate with a second male after her first mating, is an integral part of mating system evolution, sexual selection and sexual conflict. Understanding which loci are contributing to variation in this trait across populations will aid our understanding of its evolution. Whole genome sequencing was performed for iso-female lines from three different wild populations of *Drosophila pseudoobscura* that differ in the propensity of females to re-mate. The within-line propensity to re-mate has been stable for several generations within the lab strongly indicating a genetic component to the behaviour. I attempt to uncover SNPs that are consistently fixed for alternative alleles in two extremes of the phenotypic distribution, across populations. Population genomic simulation is used to estimate the proportion of fixed differences that are expected under a neutral drift scenario.

About 800 SNPs are consistently fixed for different alleles in high re-mating lines compared to low re-mating lines. This number is greater than expected by chance under a variety of different parameterisations of a population genomic simulation. Furthermore, many of these SNPs lie near genes or within regulatory regions that are

known to be involved in *Drosophila* courtship and mating behaviours and some have even been associated with re-mating rates in more classic Genome-Wide Association Studies. Given the extremely small sample size these results should be treated with some caution. Nevertheless, this study suggest that even from a relatively small sample size of isofemale lines established from wild populations it is possible to identify loci associated with a complex quantitative trait.

Author Contributions

In this chapter the extractions of genomic DNA for sequencing as well as all subsequent analysis was performed by myself. The isofemale lines were originally sampled, phenotyped and are maintained by Michelle Taylor (University of Exeter), Alison Skeats (University of Exeter), Nina Wedell (University of Exeter) and Tom Price (University of Liverpool). Other unpublished data are also used in this chapter as supporting evidence these data are indicated and were produced by Sarah Forrester (University of Liverpool).

4.1 Introduction

4.1.1 Investigating The Genetic Basis of Traits

Identifying the genetic basis of quantitative traits continues to be an important goal in biology. Moving beyond a quantitative genetic description of a trait to identify the causal loci is difficult and has led to the development of various experimental and statistical methods (Boake et al., 2002; Stapley et al., 2010; Hoban et al., 2016). Quantitative Trait Locus (QTL) mapping relies on laboratory crosses of individuals and a tracking of recombination events between markers to identify genomic regions that contribute to the phenotypic variance (Boake et al., 2002). Although many successful studies have been performed which identify QTL for traits important in adaptation (e.g. Colosimo et al., 2004; Kronforst et al., 2006) there are also difficulties (Rockman 2012; Travisano & Shaw 2012). These types of studies are laborious, typically have fairly low resolution, and there is a danger that loci with small effects are missed entirely (Rockman 2012; Travisano & Shaw 2012). As genome-wide sequencing and genotyping at large numbers of loci has become increasingly accessible, Genome-Wide

Association Studies (GWAS) have become more popular. GWAS rely on the sampling of many thousands of SNP markers throughout the genome and testing for an association between an allele and a phenotype of interest (Stapely et al., 2010; Hoban et al., 2016). GWAS have a greater genomic resolution because they can typically genotype many more markers as well as rare alleles but often require enormous sample sizes at great expense.

Innovations to experimental design in model and non-model study systems, have been accumulating (Schlötterer et al., 2014; Schneeberger 2014). For example, pooled-sequencing (pool-seq) can acquire data for large numbers of whole genomes for a fraction of the price of multiple individual whole genome sequence (Schlötterer et al., 2014). While some information is lost (e.g. haplotypes) this approach has been very successful in comparing populations in nature (e.g. Lamichhaney et al., 2012; Bergland et al., 2014; Chen et al., 2016) and in experimental evolution studies in the lab (e.g. Burke et al., 2010; Orozco-terWengel et al., 2012; Kofler & Schlötterer 2014; Schlötterer et al., 2015). Bulked segregant analysis (BSA) is another QTL approach of leveraging the extremes of the distribution of a quantitative trait in order to identify the loci underlying variation (Magwene et al., 2011). Individuals are quantified for a trait of interest and the upper and lower tails of the distribution are selected for sequencing. Regions which have no effect on the trait should have equal allele frequencies while regions with loci influencing the trait of interest should show differences in the allele frequency (Magwene et al., 2011). An extension of the concept of using extreme phenotypes has also been extended to GWAS, Extreme-phenotype GWAS (XP-GWAS; Yang et al., 2015). Causative alleles are enriched by using pools of extreme phenotypes and greater resolution is achieved by using more markers (e.g. by using SNP-chips, or whole-genome sequencing; Yang et al., 2015).

Finally, isofemale or otherwise inbred lines and induced mutant lines are a staple of research in *Drosophila* and other systems (e.g. Mackay et al., 2012; Huang et al., 2014). In isofemale lines individuals are maintained in inbred lines from a single descendant mother for several generations. Insofar as these lines then vary in phenotypes of interest, they will have captured the genetic variants underlying the trait of interest. Many such collections of lines exist as a public resource for several species including *Arabidopsis thaliana* (The 1001 Genomes Consortium 2016), *Drosophila*

melanogaster (Mackay et al., 2012; Huang et al., 2014), and humans (The 1000 Genomes Project Consortium 2015). However, these are typically from a collection of inbred lines from around the world (e.g. The 1001 Genomes Consortium 2016) or single populations (e.g. Mackay et al., 2012; Huang et al., 2014). Alternatively, researchers can set up their own reference panels for their system and population of interest. GWA or QTL studies can be done directly on these inbred lines (e.g. Montgomery et al., 2014; Ivanov et al., 2015; Gaertner et al., 2015). Alternatively, backcrosses can isolate causal regions for particular traits by repeatedly crossing offspring expressing a phenotype of interest with a parent or parent line which does not express the phenotype. This will introgress causal loci into a new genomic background and only those individuals for which the causal loci have been introgressed will express the phenotype (e.g. Schneeberger et al., 2009; Arif et al., 2013; Tanaka et al., 2015). The offspring can be sequenced and compared to parent genotypes to uncover regions showing greater similarity to the mutant parent (Schneeberger 2009; Schneeberger 2014).

The methods and approaches for identifying genetic associations with traits of interest have diversified with increasing access to the technology. Although costs of sequencing has fallen dramatically it is still high for many researchers and some methods are not feasible in many non-model species. Methods that can reliably identify associated loci while reducing the sample sizes and sequencing effort (such as pool-seq, BSA, and XP-GWAS) are particularly useful for non-model systems. Clearly there is still room for the development of novel experimental designs that can aid in the discovery of loci important in phenotypic differences.

4.1.2 Polyandry

Polyandry is defined as mating systems where females will mate with two or more different males (Boulton & Shuker 2013). Polyandry is thus characterised largely by the female re-mating rate. Originally thought to be rare, evidence has mounted that suggests polyandry is both taxonomically widespread (Taylor et al., 2014) and that it forms an important component of the evolution of mating systems (Pizzari & Wedell 2013; Snook 2015). Several hypotheses exist as to why female re-mating behaviour would be advantageous. A female can receive direct benefits, such as nutritious ejaculate components or nuptial gifts, which lead to increased survival or higher

fecundity (Arnqvist & Nilsson 2000) which makes it advantageous for her to mate with multiple males. Alternatively, females can gain indirect benefits by mating with males that will provide her with attractive or genetically superior offspring (Slatyer et al., 2012). Regardless of the initial conditions that favour polyandry it is widely accepted that female re-mating rates have important consequences for the evolution of mating systems.

Evidence from *D. pseudoobscura* suggests that one benefit of polyandry may be "bet-hedging" gains for females. Multiple mating by females reduce the chances of her offspring being sired by males who carry a costly male-gamete-killing selfish genetic element *Sex Ratio* (SR). The SR locus is associated with inversions on the X-chromosome (Sturtevant & Dobzhansky 1936; Beckenbach 1991; 1996). Males carrying the SR locus seem to produce few Y chromosome carrying sperm (Polycansky & Ellison 1970) resulting in all female broods when a female mates only with an SR male (Beckenbach et al., 1981). Population level rates of polyandry covary with latitude in the same way as the frequency of SR (Price et al., 2014). Although the high rates of female re-mating in populations in which SR is rare or absent suggest that other factors are involved in the maintenance of polyandry in this species (Price et al., 2014). Meanwhile, in experimental evolution studies the rates of polyandry increase after only 10 generation in treatments that contain SR while they remain the same in control treatments (Price et al., 2008). In these experimental evolution lines it also seems that changes to the female re-mating rate has selected for males that are better at preventing female re-mating. After 13 generations male ability to prevent female re-mating was higher in experimental treatment lines where SR was present and female re-mating had increased (Price et al., 2010). This suggests that increases in female re-mating rates results in greater conflicts over the female re-mating rate between males and females, which in turn produces selection on males to control female re-mating behaviour (Price et al., 2010). Males can accomplish an amount of control over female re-mating rates through, for example, aggressive courting or passing seminal fluid compounds which alter female behaviour (Arnqvist and Rowe 2005), although the evidence points to female re-mating being largely controlled by the female in *D. pseudoobscura*, with some manipulation by males (Price et al., 2010).

Evidence suggests that female re-mating rates in *D. pseudoobscura* have a

strong genetic component. First, female re-mating rates of wild females, estimated from the number of sires represented among the offspring, are highly correlated with the latency to re-mate among her daughters in the laboratory (Price et al., 2011). At the same time, grand-daughters of females caught in populations where re-mating rates are high tend to have higher re-mating rates, and shorter latencies to re-mate, than grand-daughters of females caught in populations with low re-mating rates (Price et al., 2014). In addition, the female re-mating rate can evolve rapidly in experimental evolution studies (Price et al., 2008).

A complete picture of the causes and consequences of female re-mating rates in *D. pseudoobscura* will require an understanding of the genetic basis of the trait. Questions about the genomics of female re-mating rate in *D. pseudoobscura* include; Is there any evidence that loci associated with higher re-mating rates are linked to the SR region? Which genes are involved in producing the variation in re-mating rate across populations? Other questions concern the great number of inversion polymorphisms in *D. pseudoobscura.* These inversions are known to occur at different frequencies across populations (Dobzhansky & Sturtevant 1937; Schaeffer et al., 2003). Inversions are increasingly being recognised as potentially important components of adaptation in clinally varying phenotypes (Kirkpatrick & Barton 2006; Hoffmann & Riesber 2008). Are loci most strongly associated with female re-mating rates within or near the common inversion regions on chromosome three?

Most work on the genetics of female re-mating has been done in *D. melanogaster* (e.g. Swanson et al., 2004; McGraw et al., 2004; 2008; MacKay et al., 2005; Giardina et al., 2011; Giardina 2015). Some candidate genes that have been associated with female re-mating rate are known. For example, olfactory receptor genes and odorant binding proteins are known to be upregulated in females as a response to mating (McGraw et al., 2004; 2008) and are associated with re-mating among female lines that vary in re-mating rate (Giardina 2011). Also implicated are many genes involved in the seminal fluid cocktail passed by the males (Ram & Wolfner 2007). Perhaps the most well studied is “sex peptide,” a male accessory gland protein which interacts with female receptors in the reproductive tract and induces a post-mating response in females of many *Drosophila* species (Chapman et al., 2003; Yapici et al., 2008; Tsuda et al., 2015). Part of this response is a reduced willingness to mate (Ram

& Wolfner 2007). Other accessory gland proteins may also have similar functions and are therefore prime candidates for genes affecting female re-mating rates. These genes are known to evolve rapidly in *Drosophila* (Haerty et al., 2007), function differently in different lineages (Tsuda et al., 2015) and some have undergone duplications or losses in different lineages (Tsuda & Aigaki 2016).

In this study I attempt to identify markers which have an assocation with re-mating rates. I take a novel approach which takes advantage of established isofemale lines that differ in their propensity to re-mate (figure 4.1). I sample lines from the tails of the distribution and identify fixed differences between pairs of lines that differ in their re-mating rates (figure 4.1). This can be thought of as analogous to three replicate BSAs (or XP-GWAS; see above) with replication being across multiple populations rather than within the extreme phenotype pools. Any single pairwise comparison will show some fixed differences between lines, many of which will be by chance. However, combining multiple pairwise comparisons from different populations (with different demographic and evolutionary histories) to identify only sites with consistently fixed differences should reduce spurious chance differences (figure 4.1). This should be especially true for species with large natural effective population sizes and genomes shaped by many rounds of recombination. These populations will have fewer chance associations between markers and phenotypes caused by large haplotype blocks because they are broken down by recombination. In this chapter the number of fixations found in three replicate pairs of lines from different populations are compared to expectations from population genetic simulations. Finally, the chromosomal positions of fixed differences and the linked genes or regulatory regions are also identified and their functional significance in the context of female re-mating is discussed.

4.2 Methods

4.2.1 Sample Collection

Samples were collected from three sites; Show Low, Arizona (34º 07' 3"N, 110º 07' 37"W); Lewistown, Montana (47º 04' 47"N, 109º 16' 53"W); and Shaver Lake (37° 8' 50.64" N, 119° 18' 6.336" W) (Price et al., 2014). Isofemale lines were set up from each population and inbred for ~50 generations in the lab. Female re-mating rate was

determined by phenotypic scoring in 2013. Virgin females were collected and stored in single sex groups of 10 individuals. At three days old, females are moved to individual vials. At 4 days old, each female was presented with a four-day old stock male and mated. Females that did not mate at this stage were discarded. Males were removed and discarded. At 8 days old each female was presented with a second four-day old stock male, and observed for two hours for any re-matings. The re-mating rate was estimated as the proportion of females that will re-mate. In total 6 isolines, two per population, were chosen from the extremes of the distribution of female re-mating rates. (figure 4.2). Summary statistics for the populations and isofemale lines are shown in table 4.1. Iso-female lines are not perfect matches in the rates of female re-mating (table 4.1) due to some lines being unavailable at the time of sequencing.

4.2.2 Sequencing and Mapping

Sequencing was carried out at the NBAF sequencing facility at the Center for Genomic Research (CGR) at the University of Liverpool. Samples were sequenced using a "pool-seq" approach (Schlötterer et al., 2014). For each isoline, 40 females were pooled and DNA extracted using a standard phenol-chloroform extraction protocol (see Appendix A). Four libraries were run on a single Illumina HiSeq lane and sequenced to ~40x coverage. Empirical coverage statistics and the number of reads generated as well as quality metrics are shown in table 4.2.

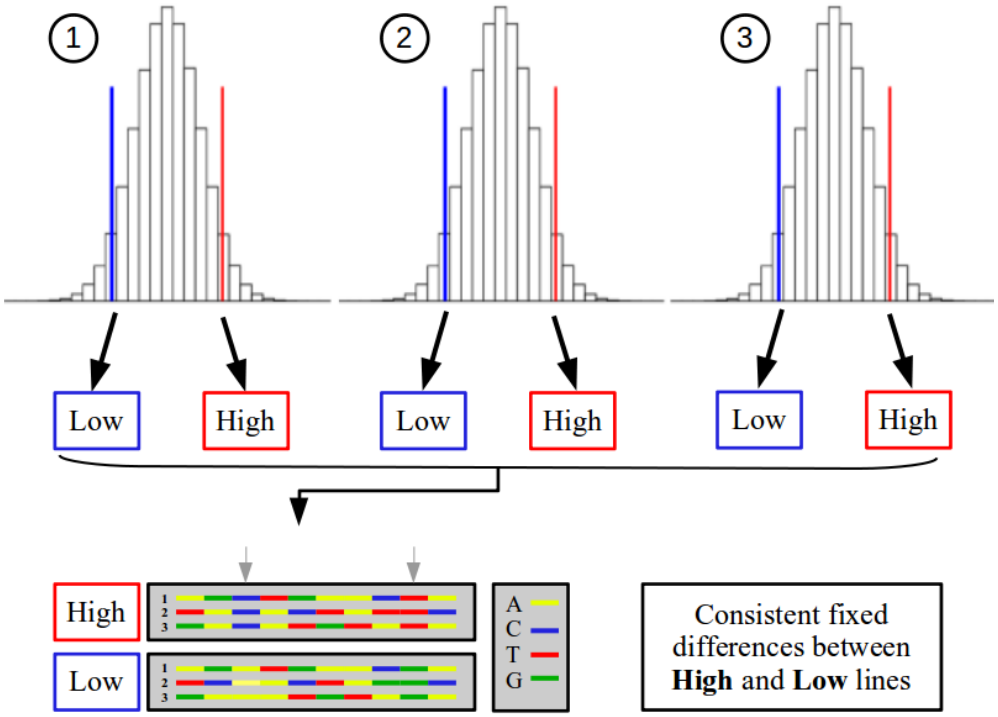


Figure 4.1. Cartoon representation of the experimental set-up. Hypothetical isofemale lines from different populations (marked "1", "2", and "3") show variation in female re-mating rates. Lines from the tails of these distributions represent "High" and "Low" re-mating rates. Pool-seq data allows the identification of SNPs that show fixed differences (grey arrows) between High and Low lines across all pairwise comparisons.

Further quality control by trimming and filtering low quality reads was performed using Trimmomatic v. 0.32 (Bolger et al., 2014). Reads were clipped if the base quality fell below Q = 20 and reads shorter than 20 bp were discarded. BWA mem (Li & Durbin, 2009; Li 2013) was used to map reads to the *D. pseudoobscura* reference genome (release 3.1, Februrary 2013) obtained from FlyBase (dos Santos et al., 2014). Duplicate reads were removed using samtools v. 1.2 (Li et al., 2009) and re-alignment around indels was carried out in GATK v. 3.3 (McKenna et al., 2010; DePristo et al., 2011). Bedtools v. 2.22.1 (Quinlan & Hall, 2010) was used to calculate various genome-wide statistics (e.g. coverage) throughout the genome. Summary statistics of the mapping step are shown in table 4.2. SNPs and allele frequencies were called with samtools v. 1.2 (Li et al., 2009) and PoPoolation2 (Kofler et al., 2011).

Due to the incompleteness of the *D. pseudoobscura* genome chromosome 4 and the X chromosome arms are split into 4 and 8 groups respectively. Meanwhile, several unmapped and fragmented scaffolds are present in the genome (17% of the genome). Unless otherwise stated, the subsequent analyses are performed on the chromosomal groups and the unknown or unplaced scaffolds ignored. Coverage across samples is fairly consistent (table 4.2), SHAA10 and SHAC1 samples have higher average coverage. To avoid any counfounding effects of large differences in coverage the .bam files for SHAA10 and SHAC1 are sub-sampled to contain 47 million alignments, corresponding to the mean across all other samples. Only biallelic SNPs with a coverage greater than 17 and lower than 59 (corresponding to the 10th and 90th percentiles of the aggregate coverage distribution respectively; figure 4.3) are considered for further analysis. If any sample does not meet these requirements the SNP is excluded. This left 3,709,701 SNPs for further analysis.

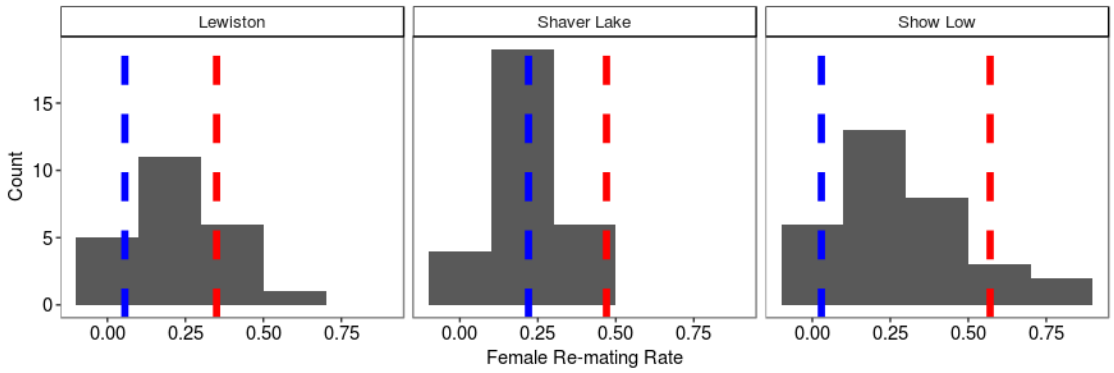


Figure 4.2. The distribution of female re-mating rates within each population. Red (high re-mating) and blue (low re-mating) vertical lines show the positions in the distributions of those lines chosen for sequencing (see table 4.1).

4.2.3 Candidate Regions

I compare my results in this chapter to results from other pilot work into the genetics of re-mating rates in *Drosophila pseudoobscura*. Work by Sarah Forrester in Tom Price's group at the University of Liverpool recently conducted a breeding experiment and sequencing designed to isolate regions that affect female re-mating rates. Briefly, isofemale lines which showed high rates of female re-mating rate were repeatedly backcrossed to a monandrous (low re-mating rate) line. Offspring from these

crosses which were polyandrous were again backcrossed to the monandrous line. After 28 generations the offspring and the parental lines were pool-sequenced and aligned to the reference genome. SHOREmap (v.0.7.0; Schneeberger et al., 2009) was then used to isolate variants which are conserved in the parental line with the same phenotype (high re-mating rates). This analysis identified 7 regions which are associated with polyandry, i.e. they are consistently shared among polyandrous offspring and the polyandrous parent line.

4.2.4 Identifying Candidate SNPs

To identify SNP variants associated with female re-mating rate I first identified all SNPs that were consistently fixed for alternative alleles in high and low re-mating rate lines (hereafter "fixed SNPs"). Pairwise comparisons were performed between lines that come from the same population, thus there were three pairwise comparisons. Because the isofemale lines are quite inbred and the sample size was relatively small (three high and three low re-mating rate lines), it is possible that many of these fixed differences might be due to chance. I used simulation to obtain an empirical null distribution of the number of consistently fixed differences expected by chance. Thus, I test whether there are more fixed differences than expected by chance.

One simulation approach considers an ancestral population at mutation-drift equilibrium in which the distribution of allele frequencies is described by a beta-binomial distribution B(α, β) (Charlesworth & Charlesworth 2008), where the α and β shape parameters which describe the distribution are given by:

$$\alpha = 4N_e u;$$

and,

$$\beta = 4N_e v$$

where N_e is the effective population size and u and v are the mutation and back-mutation rates respectively (Charlesworth & Charlesworth 2008).

Table 4.1. Summary statistics for populations and isofemale lines used in the current study. The population level re-mating rate is estimated from the proportion of clutches that show mutliple paternity in Price et al., (2014).

Population	***Isoline (Sample ID)***	***Isofemale line re-mating rate (%)***	***Population re-mating rate (%)***
Show Low	SLOB7	2.9	52
	SLOC9	57	
Lewistown	LEW17	5.6	92
	LEW23	35	
Shaver Lake	SHAA10	22	22
	SHAC1	47	

Isofemale lines can be sampled from this ancestral population by drawing allele frequencies from this distribution and assuming that allele frequencies are under Hardy-Weinberg equilibrium;

$$p^2 + 2pq + q^2 = 1$$

to determine the genotype probabilities of each female sampled. Isofemale lines are then taken to have an allele frequency of 1, 0, or 0.5 for the two homozygous genotypes and the heterozygous genotypes respectively. Then the proportion of times the allele frequency difference between pairs of isofemale lines is 1 across all *n* pairs of isofemale lines can be computed to derive a theoretical neutral distribution of such fixations.

To fully parameterise this simulation, values of N_e were obtained from the literature. Several estimates of N_e for *D. pseudoobscura* have been reported. Noor et al., (2000) estimate between 141,000 and 512,000 from microsatellite data while Jensen & Bachtrog (2011) give an estimate of 4.5×10^6 from genome-wide SNP data. Meanwhile, estimates for species with similar distributions range from 2×10^6 (*Heliconius melpomene*; Keightley et al., 2015) to 1.4×10^6 (*D. melanogaster*; Keightley et al., 2014). The simulations were thus run over the range of N_e from 1×10^6 to 4×10^6. Estimates of mutation rates also vary. However, for species similar to *D. pseudoobscura* the mutation

rate is in the range $1x10^{-9}$ to $8.4x10^{-9}$ in *D. melanogaster* (Haag-Liautard et al., 2007; Keightley et al., 2014) and $2.9x10^{-9}$ in *H. melpomene* (Keightley et al., 2015). Simulations were run over a range of mutation rates from $1x10^{-9}$ to $8x10^{-9}$. The simulations were run considering the range of $n = 2$ to 10 pairs of isofemale lines. 100 simulations were run for each value of n and the proportion of consistently fixed differences calculated from 10,000 drawn SNPs or genotypes from the distribution describing the ancestral population. These simulations should be conservative because in reality even more variation (and therefore fewer fixed differences) are expected within isofemales lines. Females from the field often will produce offspring using sperm from multiple males. Female multiple mating should result in more genetic variation among the offspring than is simulated here.

A second approach to assessing the expected number of consistently fixed differences is similar to a bootstrapping approach. Instead of a hypothetical ancestral allele frequency distribution this uses the empirical distribution of allele frequencies in the pool-seq data from each isofemale line. Because the distributions were very similar across all samples (figure 4.4 A), the aggregate distribution (summing counts in each frequency bin across isofemale line samples) was used (figure 4.4 B). Allele frequencies (p) were drawn from this empirical distribution. In this case genotypes were not assumed to be at HWE and were instead made by simply drawing two alleles at random from a binomial distribution where the probabilties were given by p. Isofemale lines are then given allele frequencies of 1, 0, or 0.5 as above. Finally, the proportion of times consistent differences of 1 (i.e. fixed differences) across all pairs of isofemale lines are seen is computed. The proportion of consistent fixations was calculated from 10,000, 100,000, and 1,000,000 SNPs drawn from the empirical distribution across $n = 3$ pairs of isofemale lines. A distribution of the proportion of consistent fixations was simulated from 100 runs for each number of SNPs.

This bootstrapping approach should be conservative because there is very little variation in allele frequencies in the empirical distribution with most SNPs being fixed for either the major or minor allele, thus there is a higher probability of individuals being homozygotes but roughly an equal probability of individuals being homozygous for the major or minor allele.

4.2.5 Funtional Aanalysis

For the identified SNPs a functional analysis was carried out by Gene Ontology (GO) term and phenotypic class enrichment analysis. These analyses rely on GO term and phenotypic associations with annotated genes in *D. melanogaster*. Thus, *D. melanogaster* GO terms were downloaded via FuncAssociate (v2.0; Berriz et al., 2009). The *D. pseudoobscura* annotated genes were converted to *D. melanogaster* orthologs, where duplicates were found they were re-labelled in the annotation and the duplicate ID was added to the GO term dataset. GO term enrichment analysis was performed for each SNP in GOwinda (v1.12; Kofler & Schlötterer 2012). The SNP was considered "genic" if it occurred within 1Mb up- or down-stream of an annotated gene. GOwinda was run with default paramaters and the empirical null distribution of gene abundance distribution obtained by 1,000,000 simulations.

Phenotype enrichment analysis was performed with DroPhEA (Weng & Liao 2011). First SNPs were associated with a gene by identifying the closest gene within 1Mb to each fixed SNP. The set of all unique genes was submitted to DroPhEA to test for an association with any phenotypic classes. A distance of up to 1Mb in GO term and phenotype enrichment analysis is justified on the basis that regulatory regions are frequently mapped to distances of ~5kb (Werner et al., 2010), ~20kb (Chan et al., 2010), and up to 1 Mb up- or downstream from a target gene (e.g. Maston et al., 2006; Pennachio et al., 2013).

Finally, a transcription factor (TF) motif enrichment analysis was performed with the AME routine from the MEME package (v. 4.10.2; Bailey et al., 2009). This tool takes a set of short DNA sequences and compares them to a database of known TF binding motifs to determine if any are overrepresented among the sequences. The sequence extending 30bp up- and down- stream of each fixed SNP was extracted from the genome. This region is large enough to accommodate even the larger TF binding motifs but small enough that the focal SNP could conceivably be in within the active region of the motif. An archive of scripts and a description of the pipeline can be found at https://github.com/RAWWiberg/ThCh4.

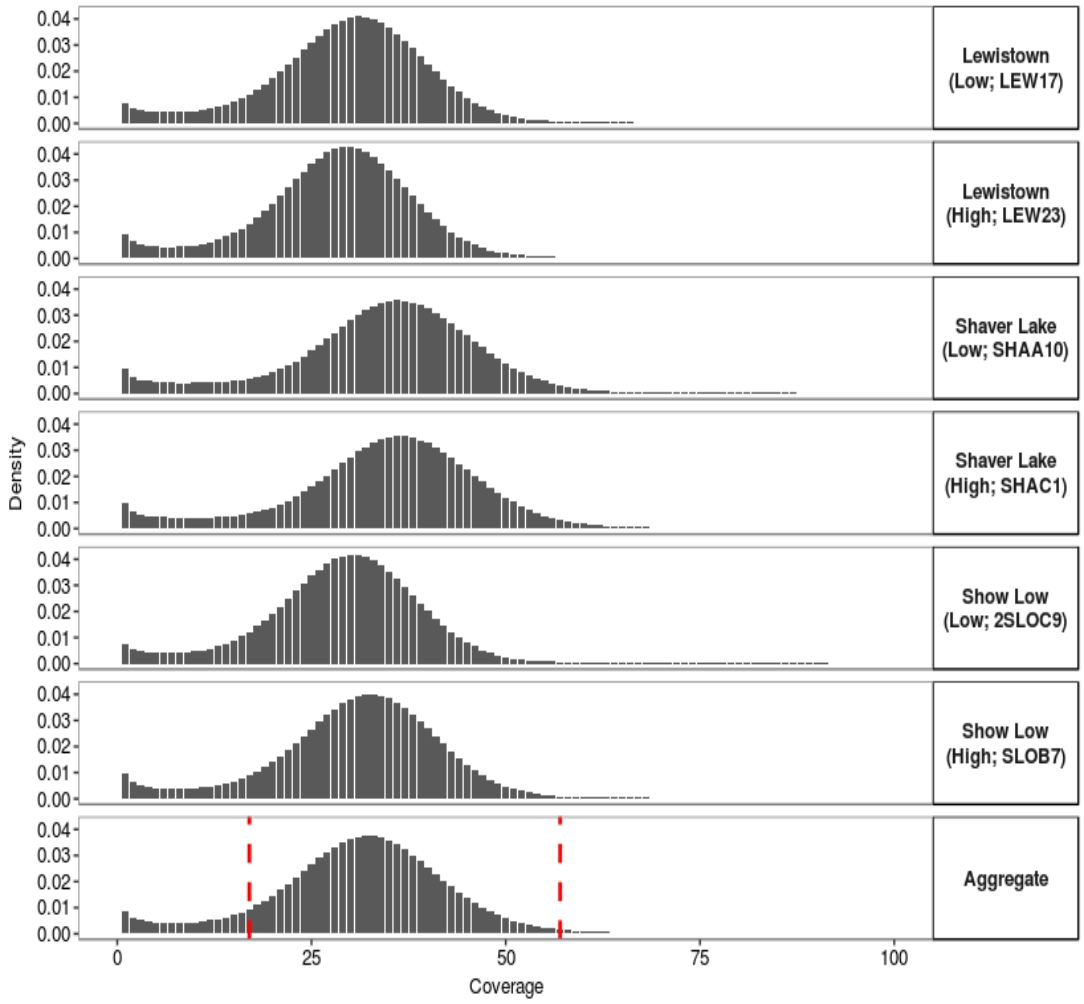


Figure 4.3. Coverage distributions across samples. Panels and panel labels refer to the isofemale lines in Table 4.1 and give the qualitative levels of female re-mating rates in each line. The "aggregate" distribution is obtained by summing counts in each coverage bin across samples. Red vertical lines give the 10th and 90th percentiles.

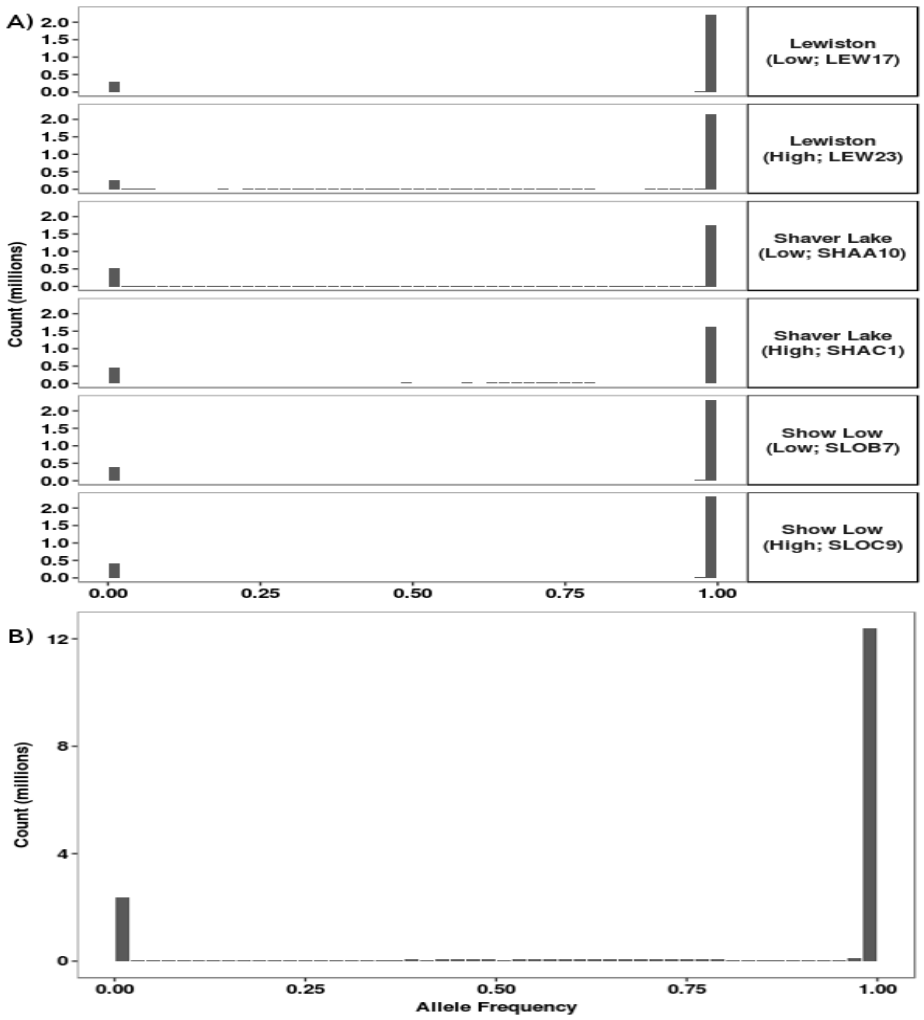
4.3 Results

4.3.1 Mapping

Mapping results are given in table 4.2 and figure 4.3. Sub-sampling of the SLOB7, SHAC1 and SHAA10 samples equalises the coverage across samples (table 4.2). The majority of reads were mapped unambiguously and properly paired with forward and reverse reads mapping to the same scaffold (table 4.2).

Table 4.2. Summary statistics of sequencing, quality filtering and mapping steps for each sample. Figures before removal of duplicate reads, indel re-alignment, and sub-sampling of SLOB7, SHAC1, and SHAA10 samples are given in square brackets. The coverage distributions are shown graphically in figure 4.1.

Isoline (Sample ID)	***Number of reads***	***% mapped (% proper pairs)***	***Mean coverage***
SLOC9	~51.02 m	100 (97.92)	36.52x
	[~51 m]	[100 (97.94)]	[37.08x]
SLOB7	~78.09 m	100 (97.89)	34.49x
	[~79 m]	[100 (97.92)]	[55.51x]
LEW17	~49.07 m	100 (97.93)	35.16x
	[~49 m]	[100 (97.95)]	[35.71x]
LEW23	~47.0 m	100 (97.94)	33.83x
	[~47 m]	[100 (97.95)]	[34.83x]
SHAC1	~47 m	100 (95.3)	42.38x
	[~151.2 m]	[100 (95.30)]	[124.5x]
SHAA10	~47 m	100 (96.30)	42.11x
	[~109.6 m]	[100 (96.30)]	[92.0x]



a) **Figure 4.4.** Allele frequency distributions of the overall major allele in all populations for **A)** all populations seprately, and **B)** summing counts from each bin across the populations into an aggregate distribution. Panel titles in **A)** give the population name, the level of female re-mating and the sample ID.

4.3.2 Indentifying Candidate SNPs

In total, 3,709,701 SNPs were called and passed quality control. Out of these, 816 SNPs (0.022%) are consistently fixed for the same alleles in the high re-mating lines when compared to the low re-mating rate lines across all pairwise comparisons (fixed SNPs). The genomic locations of the fixed SNPs are given in figure 4.5. Out of 816 fixed SNPs, 24 (3%) lie within the candidate regions identified by Sarah Forrester (see *4.2 Methods*). The largest concentrations of fixed SNPs is on the 4th chromosome

(41% of all fixed SNPs), followed by both arms of the X chromosome (XR = 28%; XL = 12% of all fixed SNPs). A spearman rank correlation finds no assocation between chromosome length and the number of fixed SNPs if chromosome 4 and the X chromosome are considered as complete chromosomes rather than fragmented regions (rho = 0.5, S = 10, p = 0.45). However, if the correlation is performed keeping these chromosome fragments separate there is a positive correlation between the length of the chromosomal region and the number of fixed SNPs (rho = 0.76, S = 134.18, $p < 0.001$). Of course, these results should be treated with some caution because the individual chromosomal regions are not completely independent.

Simulations from a population at mutation-drift equilibrium suggest that 816 SNPs is more than would be expected by chance (figure 4.6). As expected, the proportion of consistently fixed SNPs increases with N_e and decreases with increasing numbers of sampled isofemale lines (figure 4.6). Differences in mutation rates do not seem to have much of an effect on the number of fixed SNPs except at higher N_e. For $n = 3$, The range of the 95th percentile of the distribution of expected proportions of fixed differences is between 0% and 0.01% (figure 4.6). This suggests that 0.022% is a significantly greater proportion of fixed differences than expected by chance. Only at very high effective population sizes (3 million and 4 million) and low sample sizes ($n = 2$), are the 95th percentiles higher than 0.022% (between 0.01% and 0.07%). In contrast, bootstrap sampling of SNPs and allele frequencies from the empirical distribution suggests that the proportion of consistently fixed differences expected by chance is around 1.8% regardless of the number of SNPs sampled (10,000, 100,000, or 1,000,000; figure 4.7).

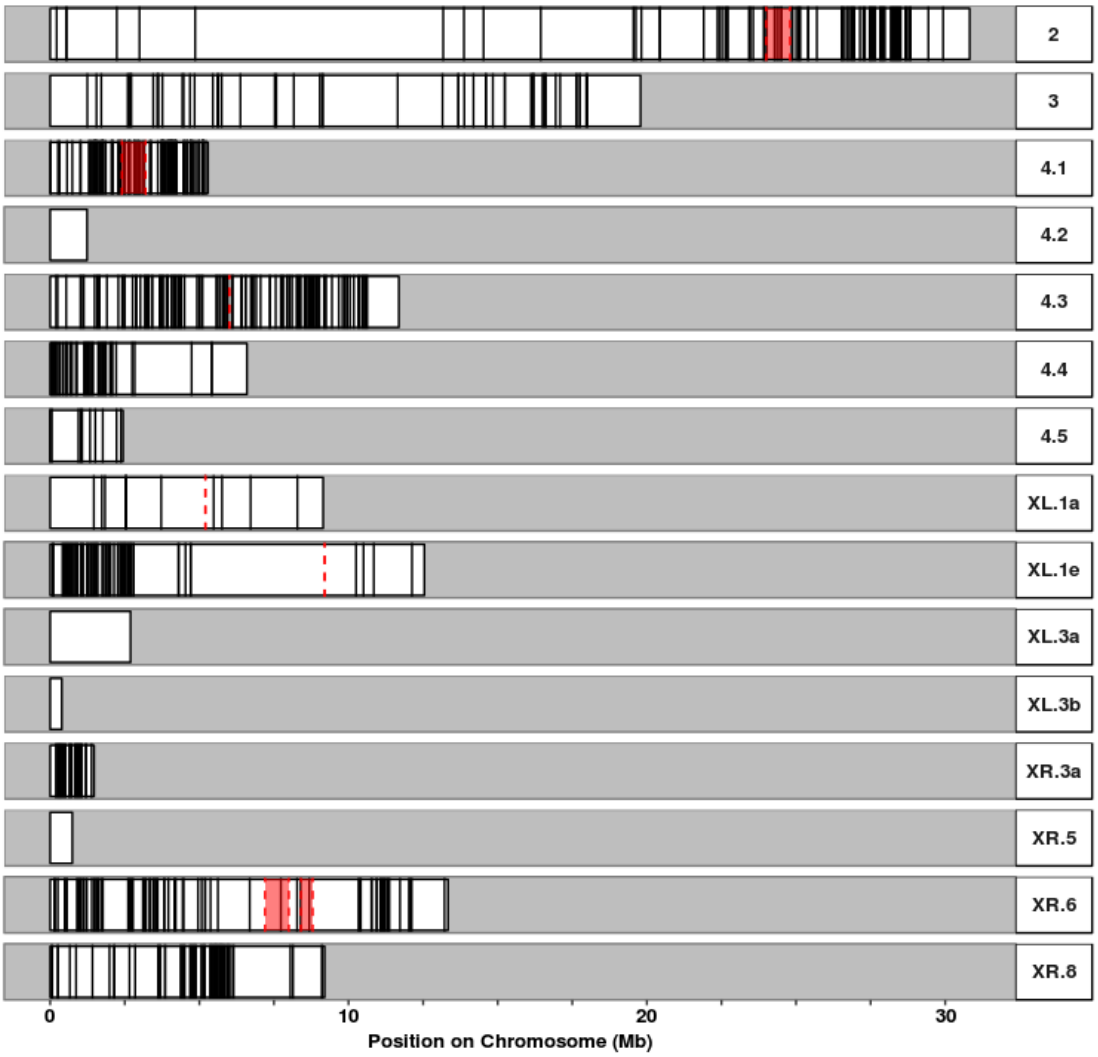


Figure 4.5. Ideogram showing the chromosomal locations of 816 SNPs with consistently fixed differences between isofemale lines. Regions delineated by shaded red and vertical red, dashed lines give the locations identified by Sarah Forrester (see *4.2 Methods*). Panel titles give the chromosome names. For chromosome 4 and the X chromosome arms the chromosomes are split into groups (see *4.2 Methods*).

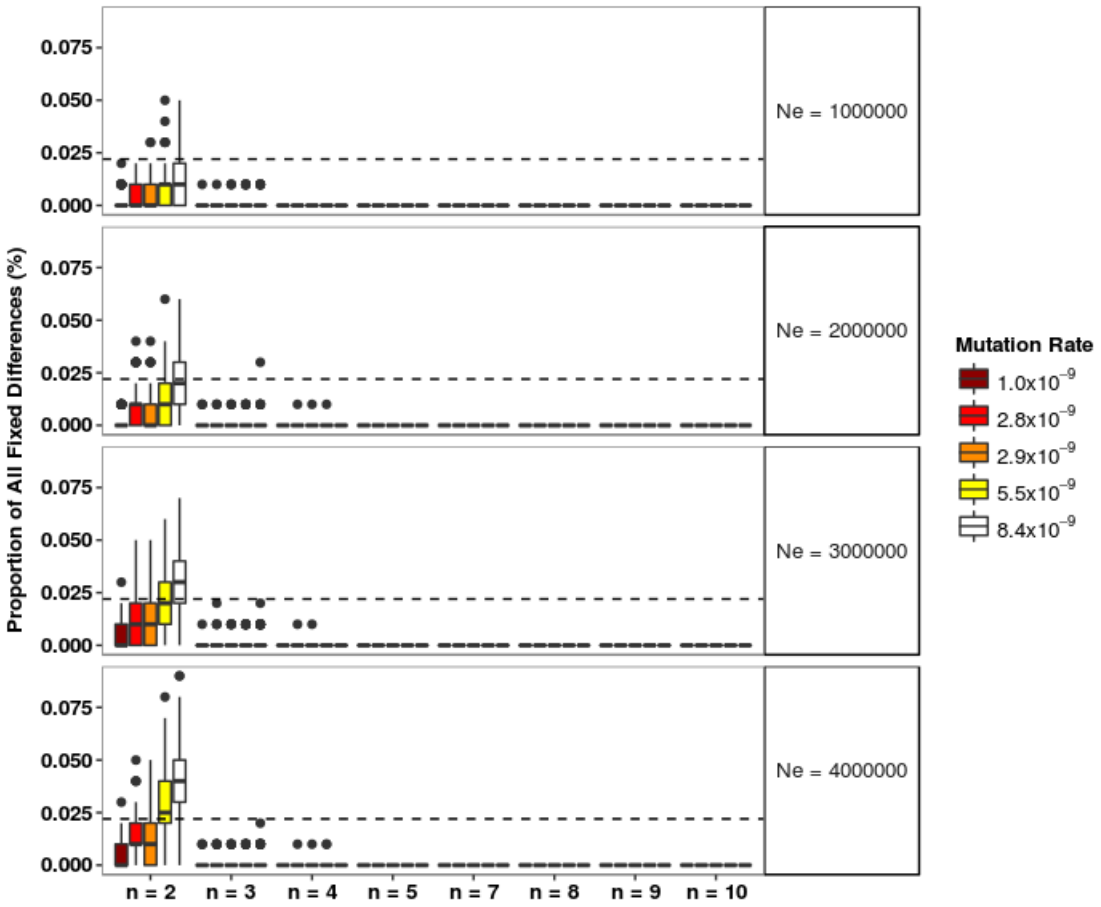


Figure 4.6. The proporion of SNPs that are consistently fixed between *n* pairs of isofemale lines. Data are from population genetic simulations with different parameter combinations. Each panel is for a different value of N_e. The *x*-axis shows different numbers of isofemale lines pairs (*n*). The horizontal dashed line gives the results seen in the pool-seq data (0.022%).

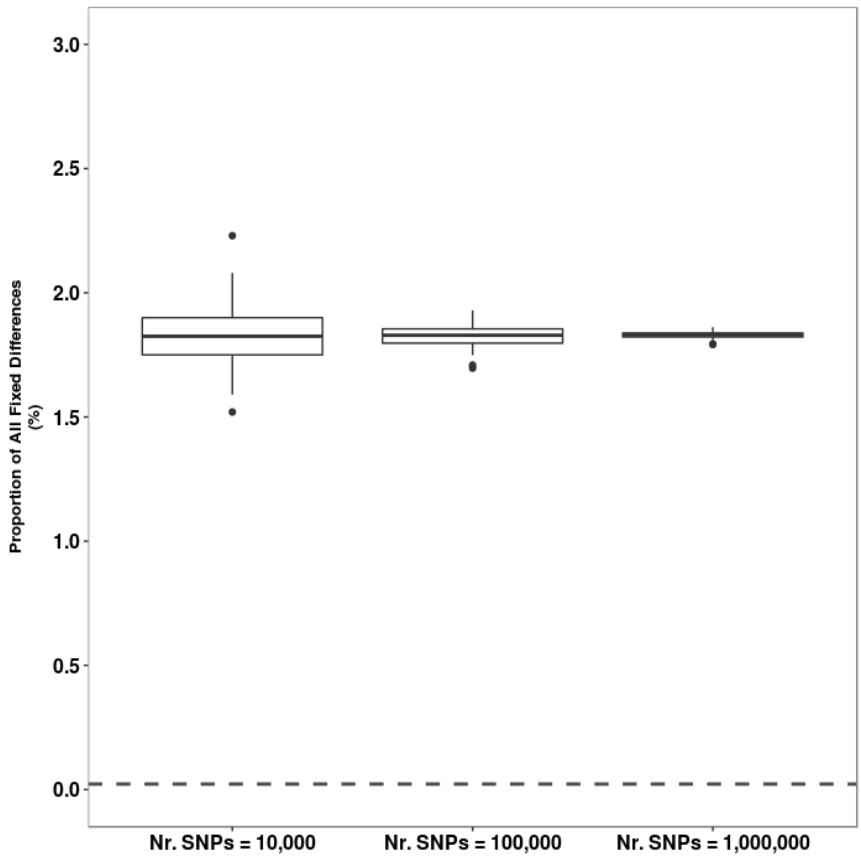


Figure 4.7. The expected proportion of consistently fixed differences between *n* = 3 pairs of isofemale lines. The proportions are based on samples of 10,000, 100,000, or 1,000,000 SNPs, variation in the estimates comes from 100 bootstraps. The horizontal dashed line gives the results seen in the pool-seq data (0.022%).

4.3.3 Functional Analysis

GO term enrichment analysis finds no enrichment of GO terms for genes within 1kb of all fixed SNPs (all $p > 0.05$ after correction for multiple testing). Neither is there an enrichment of any GO terms for genes within 1kb of SNPs within the candidate regions in figure 4.5. These results hold if all genes within 1Mb of fixed SNPs are considered. In a separate functional analysis, all 816 SNPs from the set of all fixed SNPs discovered lie within 1Mb of 498 unique gene regions. Genes within 1Mb of a fixed SNP are significantly enriched for genes in various phenotypic classes including "behaviour defective" (bonferroni adjusted p-value = 0.009), "neuroanatomy defective"

($p < 0.001$), "planar polarity defective" ($p = 0.002$), "cell polarity defective" ($p = 0.005$). Of particular note are the genes within the "behaviour defective" class. several of these genes; *longitudinals lacking (lola)*, *spineless (ss)*, *garnet (g)*, *CG33158*, *reaper (rpr)*, *odorant receptor 67 d (Or67d)*, *syntrophin-like 1 (Syn1*), have mutant phenotypes which cause defective mating behaviours. Genes within 1Mb of the subset of fixed SNPs that also occur within the candidate regions are not enriched for any phenotypic classes. In total four odorant receptor genes occur within 1Mb of a fixed SNP (*Or33c*, *Or22c*, *Or67d*, and *Or65c*). Table 4.3 shows genes that are observed within 1Mb of a fixed SNP in this study and are also either within the phenotypic class "mating behaviour defective" or known from other studies in female *Drosophila* re-mating behaviour.

TF binding site motifs for a variety of TFs are enriched around fixed SNPs. In particular, motifs for *lola* are enriched among the regions around fixed SNPs (11 motifs with adjusted p-values < 0.05). Motifs for *disconnected* (*disco*; two motifs with adjusted p-values < 0.05) are also overrepresented among the sequences near fixed SNPs. Motifs for *fruitless* (*fru*; 1 motif with ajusted p-values < 0.05) and *doublesex* (*dsx*: two motifs with adjusted p-values < 0.05) which are important in sex determination and the development of sexually dimorphic phenotypes.

4.4 Discussion

4.4.1 Identifying Candidate SNPs

Variation in insect mating systems has profound consequences for the evolution of particular traits. Insect mating systems are characterised in large part by the female re-mating rate which modulates the strength of sexual selection and sexual conflict. Therefore, understanding the genetics of female re-mating behaviour is fundamental to understanding the evolution of insect mating systems. In this study I identified a number of genomic markers that are consistently fixed for alternative alleles in pairwise comparisons of high and low re-mating isofemale lines from different populations. These SNPs represent strong candidates for loci associated with female re-mating.

Table 4.3. FlyBase IDs, gene names, references, and notes for genes of interest that lie within 1Mb of a fixed SNP.

FlyBase ID	***Gene name***	***Note***	***Comment***
FBgn005630	*longitudinals lacking*	*+cf*	-
FBgn0003513	*Spineless*	*f*	-
FBgn0001087	*garnet*	*+f*	-
FBgn0053158	*CG33158*	*+f*	-
FBgn0011706	*reaper*	*f*	-
FBgn0036080	*Odorant receptor 67 d*	*f*	-
FBgn0037130	*Syntrophin-like 1*	*f*	-
FBgn0032601	*yellow-b*	*+d*	*Yellow-g h*omolog found in *d*
FBgn0034808	*CG9896*	*e*	*CG9897* in close proximity to *CG9896* found which has a strong association in *d*.
FBgn0041181	*Thioester-containing protein 3*	*+b*	*Tep4* homolog found in *b*. Also marginally significant association in *d*.
FBgn0000721	*foraging*	*+c*	-
FBgn0003165	*pumilio*	*+c*	-
FBgn0001978	*shuttle craft*	*c*	-
FBgn0002543	*roundabout 2*	*c*	*Robo1* homolog found in *c*.
FBgn0263995	*couch potato*	*+c*	-
FBgn0000045	*Actin 79 B*	*a*	*Act88F,* homolog found in *a* and *d*

a, found in McGraw et al., (2004);

b, found in Swanson et al., (2004);

c, found in Mackay et al., (2005);

d, found in McGraw et al., (2008);

e, found in Giardina et al., (2011);

f, found in phenotypic class FBcv0000387 = "behaviour defective."

+, also occur within 1kb of fixed SNPs

I use simulations and bootstrapping approaches to obtain conservative estimates

of the number of consistently fixed differences that are expected under a neutral null model and by chance. The population genetic simulations are expected to give conservative estimates. In reality, sampled females from a population will have mated with at least one male whose genotype will also be represented among the offspring (David et al., 2005; Price et al., 2011). This should increase the genetic diversity within the isofemale line and reduce the probability of finding a fixed difference with another isofemale line by chance alone. If a female has mated with more than one male the genetic diversity among her offspring will be higher still and, consequently, the probability of finding a fixed difference with another isofemale line by chance is reduced. On the other hand, more inbreeding during maintenance in the lab will tend to reduce diversity within the lines and probably increase the number of fixed differences expected by chance. Additionally, homozygous lethal alleles will never be fixed in isofemale lines thus some residual heterozygosity is always expected. The values from the simulations can be taken as a maximum expectation.

Similarly, the bootstrapping approach should be conservative because the distribution of allele frequencies from which bootstrap samples are drawn is a reflection of the variation among the isofemale lines in this study. Because these lines reflect only a subset of the variation found in wild populations the possible diversity is greatly reduced leading to more consistent fixations in the bootstrap simulations. For these reasons the expected proportion of consistent fixations estimated from population genetic simulations and the bootstrapping protocol are likely to conservatively high. The bootstrapping approach is likely too conservative because the distribution of allele frequencies is unrealistically biased toward fixed alleles which will increase the number of fixed differences observed. A better bootstrapping appraoch would consider allele frequencies from across all isofemale lines available from the source populations. A potential problem with both of the above approaches is that they do not account for linkage between SNPs. Closely linked SNPs will be in linkage disequilibrium (LD) and their allele frequencies highly correlated. Since many of the fixed SNPs in this study seem to occur in clusters (figure 4.5) this may be inflating the number of fixed SNPs above the expected number from simulations. However, the fact that pairs of lines are from completely different source populations should ensure that recombination has broken down LD between even very closely linked SNPs.

Another potential approach to producing a "null" expectation for fixed differences is to use a permutation approach. In this approach the labels for each sample could be swapped repeatedly and the number of fixed differences re-calculated. With 6 samples and 2 labels there are a limited number of combinations. One could use the "pool" of 6 samples and sample any number of pairs, two at a time, with replacement, from this pool to make up random pairs of samples. This would maintain the linkage patterns between SNPs and their allele frequencies present withing the actual iso-female lines. However, this appraoch would suffer similar limitations to the bootstrapping approach in that it limits itself to the diversity present in the iso-female lines, which is only a small representation of natural variation. Alternatively, drawing alleles as in the simulations but building haplotypes using a empirical values of LD and recombination rates from *D. pseudoobscura* might allow for more sophisticated simulations.

Nevertheless, I conclude that the population genetic simulations are a useful method for determining the expected number of fixations and that it is superior to the bootstrap approach. These simulations are based on established solutions to population genetic equations describing the behaviour of alleles under neutral drift and mutation. At the same time the simulations are parameterised with a range of mutation rates and effective population sizes that reflect the best estimates from natural populations of species like *D. pseudoobscura*. Variation in these parameters does not change the conclusions of the study. Therefore, fixed SNPs identified in this study are good candidates for assocations with variation in re-mating rate among these lines.

4.4.2 Genes and Regulatory Motifs Near Fixed SNPs

Fixed SNPs identified in this study lie in close proximity to several genes that have been implicated in female re-mating behaviour in previous studies. For example, four odorant receptor genes (*Or33c*, *Or22c*, *Or67d*, and *Or65c*) occur within 1Mb of a fixed SNP. Microarray studies of mated females show that odorant binding proteins are upregulated in response to having mated (McGraw et al., 2004; McGraw et al., 2008) sometimes even in the absence of the transfer of seminal fluid compounds or sperm (McGraw et al., 2004). These binding proteins relay information to odorant receptors which modify female post-mating behaviour, including her receptivity to further mating attmepts (Leal 2013; Tram & Wolfner 1998). Odorant binding proteins have also been

shown to change expression patterns in lines selected for slow and fast mating latency (Mackay et al., 2005). One odorant binding protein (*Obp56a*) has also been associated with female re-mating rates in an assocation study of 91 *D. melanogaster* second chromosome substitution lines (Giardina et al., 2011). This same study found an assocation with a serine endopeptidase (*CG9897*; Giardina et al., 2011). Chromosome 2 is homologous with chromosomes 3 and 4 in *D. pseudoobscura* (Richards et al., 2005).

Other interesting genes include, *yellow-g,* important in eggshell formation, which is upregulated in females after mating (McGraw et al., 2004). *Yellow-b,* a homolog of *yellow-g,* is observed among the genes near fixed SNPs in this study. Another gene *CG9897* which is found within 1Mb of a fixed SNP has a highly significant association with female re-mating rates in Giardina et al., (2011). It does not occur in the lists of genes within 1Mb of fixed SNPs because the gene *CG9896* is closer to the nearest fixed SNP. This highlights the complexity of identifying causal loci from nearby markers in regulatory regions. Genes may share regulatory regions or may be regulated from great physical distance and regulatory regions may even lie within introns of neighbouring genes of a different function (Kleinjan & van Heyningen 2005; Whitkopp & Kalay 2012).

Other notable genes implicated in this study include *Tep3,* which belongs to a group of thioester containing proteins (*Teps*) which have an endopeptidase inhibiting function and are involved in the immune response and are known to evolve rapidly in *Drosophila* (Jiggins & Kim 2006). Upregulation of immune response genes in response to courtship or mating is often noted in response to mating and courtship in females (McGraw et al., 2004; 2008; Innocenti & Morrow 2009; Giardina et al., 2011; Immonen & Ritchie 2012; but see Immonen et al., 2017). Some evidence suggests this may be an adaptive response to protect against sexually transmitted infection (Zhong et al., 2013). Another gene, *foraging* (*for*), is involved in larval foraging behaviour and long-term memory (Sokolowski 2001). It is also known that female re-mating rate, in *D. melanogaster*, is strongly related to feeding behaviour with resource starved females being less eager to re-mate as long as sperm is not limited (Harshman et al., 1988). The gene *pumilio* (*pum*) is a post-translational regulator of transcription (Gerber et al., 2006). It is implicated in the formation and maintenance of long-term memory in *Drosophila* (Dubnau et al., 2003). It is also involved in the development and

maintenance of female germ cells (Gerber et al., 2006). In experimental lines it shows changes of expression after selection for fast and slow mating rates in *D. melanogaster.* Flies with a short latency to re-mate have higher expression of *pum* (Mackay et al., 2005). The genes *shuttle craft* (*stc*), *roundabout 2* (*robo2*) and *longitudinals lacking* (*lola*) are all TFs and involved in neurogenesis (Giniger et al., 1994; Stroumbakis et al., 1996; Neumüller et al., 2011; Evans et al., 2015).

Transcription factor (TF) binding site motifs for several TFs are enriched in regions around the fixed SNPs. In particular transcription factors that are involved in sex determination (*fru,* and *dsx*) as well as neurogenesis (*lola* and *disco*) are overrepresented among all the transcription factor binding motifs. *lola* and *disco,* and to a lesser extent *fru* and *dsx*, have been shown to change expression patterns in selection lines for slow and fast mating latency (Mackay et al., 2005). All of the above TFs are known to have important roles in courtship behaviours (Sokolowski 2001). For example, *fru* is known to regulate the development of sexually dimorphic nervous systems in *Drosophila melanogaster*. Male specific mutations result in increased latency to courtship, longer time to copulation and reduced overall amounts of courting behaviour (Neville et al., 2014). The TF *lola* itself is also in close proximity to some fixed SNPs in this study (table 4.3).

The molecular mechanisms which alter female re-mating rates will depend on the ancestral mating system. If this mating system was predominantly characterised by aggressive males and females resisting male advances then selection for increased re-mating rates in females simply could be a relaxation of selection to resist male mating attempts (e.g. relaxed selection on countering male accessory gland proteins). On the other hand females may already have been benefitting from re-mating with multiple males in which case selection will act on alleles at genes that alter female mating periodicity more directly. Several of the genes identified in this study are known to be involved in rates of re-mating among males. One interpretation is that female re-mating rates are, at least in part, genetically correlated with male re-mating rates. This might come about if the oft cited conflict between males and females over the optimal re-mating rate (Arnqvist & Nilsson 2000) is relaxed in this system because higher rates of re-mating are partly adaptive among females (they avoid SR carrying males). This would relax selection on females to control expression of genes that influence mating

rates to female specific optima. On the other hand, several of these genes have also been implicated in the intrinsic variation in female re-mating rates that is not corellated with the mating rate of sons or with the transfer of sperm and seminal fluids (e.g. *Obp56a*; Giardina et al., 2011). These findings point to separate mechanisms underlying variation in female re-mating rate that are not the result of intra-genomic conflict over phenotype expression.

Given the small sample size, these results need to be treated cautiously. Scope for improvement would include the sampling and sequencing of more isofemale lines to look for consistently fixed differences and checking for an association between re-mating rate and the markers identified here. Additionally, population genomic samples could be taken from populations across the cline to ask if the SNP alleles reported in this study also vary clinally with the re-mating rate and to uncover new marker alleles that vary clinally or that covary with population level re-mating rates. This said, it is encouraging to observe results that are broadly consistent with previous work on female re-mating rates in *D. melanogaster.*

4.5 Concluding Remarks

This study attempts to identify SNP markers that are associated with the female re-mating rate in *Drosophila pseudoobscura*. Isofemale lines from different populations, that differed in their re-mating rates, were whole-genome sequenced to identify markers that are consistently fixed for alternative alleles in high and low re-mating lines. Population genetic simulations suggest that the number of fixed differences observed is greater than expected by chance. Many genes in close proximity to fixed SNPs have been directly implicated, or are related in function to implicated genes, in female mating behaviour from previous studies. This study demonstrates the feasibility of a novel experimental breeding (maintaining isofemale lines) and selective sequencing approach (extreme phenotypes) to uncover loci associated with interesting tratis. This study also identifies promising candidates for follow up work on the genomics of female re-mating rates which will give a clearer picture as to how selection has shaped this trait in natural populations of *D. pseudoobscura*.

"...to direct attention to variation within groups...I propose the word cline...a gradation in measurable characters."

- J. Huxley, 1938

Chapter 5 The relationship between genomic and environmental differentiation among populations of *Drosophila montana.*

Abstract

The study of clinal phenotypes can help us understand the forces that drive population differentiation and speciation. Traits that vary clinally give an indication toward the selective pressures faced by species in their environment. At the same time, identifying the genomic loci that show allele frequency clines can uncover loci important during population differentiation and the early stages of speciation. Inferences about the forces that shape these clines can be made from population genetic summary statistics and fitting different models of demography to the genetic data.

Drosophila montana shows clinal variation in several ecologically important phenotypes. In particular the critical day length, the photoperiod after which 50% of females enter diapause, is much shorter for northern populations. This effect is observable among isofemale lines established from the wild even after many generations of laboratory culture and Quantitative Trait Loci (QTL) have been mapped, indicating a genetic and inherited component. I take a population genomic approach to identify loci that show signatures of local adaptation. First I quantify the environmental (climatic) variation among populations of *D. montana.* Using several environmental variables it is possible to find the principle climatic axes that characterise differences among populations. Then I use recent Bayesian methods which relate the amount of genomic differentiation to the environmental differentiation between populations. I also

test quasibinomial GLMs in a scenario using a continuous rather than categorical predictor, a possibility I identified in Chapter 2.

Many SNPs show strong associations with the main axes of environmental differentiation. The closest genes to these top SNPs are also of note for their roles in relevant phenotypes. Many of them have also been identified as candidate genes in diapause and cold-tolerance behaviours from previous expression studies lending further support to their functional relevance. The regions around some of these genes show evidence of selective sweeps or background selection in all or some of the populations indicating ongoing selection at these loci. Finally, I note that quasibinomial GLMs did not perform as well as expected and should probably be used only with considerable caution in these types of analyses. Bayesian methods show much more promise.

Author Contributions

For this chapter I designed the pipeline and performed all of the analyses. Original field sampling was conducted by Anneli Hoikkala (University of Jyväskylä), Mikko Hoikkala, Jackson Jennings (University of Arkansas), Antti Miettinen, and Hannele Kauranen (University of Jyväskylä). Sample curation and DNA extractions were coordinated by Maaria Kankare (University of Jyväskylä) with Riikka Tapaninen and Johanna Kinnunen. I also use data from the as yet unpublished *D. montana* reference genome where multiple people have contributed for a full list of contributors to the genome project see the references (Parker et al., *in prep*).

5.1 Introduction

5.1.1 Population Clines

As species diversify they encounter new ecological niches which impose selection pressures. Some novel mutations or variants already present in the population will provide a selective advantage and allow a group to spread further into a new environment. If the environment in question follows a gradient, e.g. along latitudes of the globe, this process can produce a series of populations which are locally adapted to the environment and display a gradient of trait values (Huxley 1938; Haldane, 1948;

Endler 1977; Takahashi 2015). The study of such clinal populations can uncover traits that are important in ecological adaptation and speciation (Endler 1977; Takahashi 2015). Insofar as these traits are heritable, the study of clinally distributed phenotypes can also help us understand the genetic basis of adaptive traits, and how population divergence and speciation progresses in the face of homogenising gene flow (Endler 1973; Endler 1977, Kirkpatrick & Barton 1997; 2006).

With the recent expansion of Next-Generation Sequencing (NGS) methods, researchers have taken advantage of clinal variation in some species to uncover the genetic variants which underlie these important phenotypic traits. Several studies have used pooled-sequencing (Pool-seq) and methods based on identifying F_{ST} outliers to study adaptation and patterns of genome differentiation across environments. For example, Kolaczkowsi et al., (2011) sampled isofemales lines from the north and south of the well studied Australian *D. melanogaster* cline. The study used pool-seq to estimate allele frequencies and measure differentiation (F_{ST}) between populations and to identify outlier regions throughout the genome (Kolaczkowski *et al.,* 2011). The study found substantial differentiation in intergenic and non-coding regions, although many genes implicated in phenotypes that also vary clinally were also in the most differentiated regions (Kolaczkowski et al., 2011). In another study, three populations of *Anopheles gambiae* were sampled from a cline in Cameroon to compare differentiation within and outside a potentially important inversion polymorphism (Cheng et al., 2012). F_{ST} was estimated along genomic windows for populations at opposite ends of the cline to find that differentiation between population was almost entirely localised to the known inversion regions (Cheng et al., 2012). Genes near the most differentiated SNPs were enriched for similar functional categories as in Kolaczkowski et al., (2011) raising the possibility that convergent adapation to similar environmental stresses involves the same biochemichal pathways in both species (Cheng et al., 2012).

More recently studies have sought to model clinal distribution of markers more continuously in space or along environmental gradients rather than differentiation between pairs of populations. For example, a recent study sampled 6 populations in North American clines of *D. melanogaster* and *D. simulans* over several years to uncover genomic markers underlying traits that vary clinally as well as seasonally (Bergland et al., 2014; Kapun et al., 2016). These studies find that many SNPs show

consistent seasonal fluctuations in allele frequencies throughout the cline indicating a response to seasonally varying selection pressures (Bergland et al., 2014). Meanwhile , the allele frequencies at 'seasonal SNPs' in northern populations were more 'fall-like' while those in southern populations were more 'spring-like' (Bergland et al., 2014). Together these results point to adaptive differences in allele frequencies in different parts of the cline with polymorphism maintained throughout the species by seasonally varying selection. In another study some common inversions polymorphisms were shown to have a stable cline among these populations in the face of gene flow while others showed signs of seasonally fluctuating frequencies (Kapun et al., 2016). Comparing similar distributions of *D. simulans* and *D. melanogaster* Machado et al., (2015) conclude that migration and gene flow play a greater role in the overall clinality of genomic variants in *D. simulans* than in *D. melanogaster*. Nevertheless, the two species share a significant proportion of the genes associated with clinal SNPs. The authors highlight the differences in physiological tolerance to overwintering between the species and the resulting differences in migration and bottlenecks as additional drivers of differences in genomic variation between them (Machado et al., 2015).

Though the focus here is primarily on *Drosophila,* similar studies of clinally varying phenotypes and genetic alleles have also been done in other insects (e.g. Paolucci et al., 2016), plants (e.g. Chen et al., 2012; Bradbury et al., 2013), mammals (e.g. Hoekstra et al., 2004; Carneiro et al., 2013), fish (e.g. Vines et al., 2016), and many others (Endler 1973; Endler 1977; Takahashi 2015; Adrion et al., 2015). Clearly, the study of environmental clines can shed light on the process of adaptation as well as on the differences in these processes between species that contribute to biological diversity.

5.1.2 Drosophila montana

Drosophila montana is a species in the *Drosophila virilis* species group which has spread around the northern hemisphere. During this spread it has adapted to many different latitudes and is one of the most northernly distributed species of *Drosophila* species (Throckmorton 1982). The distribution of the species imposes particular sources of selection from the environment that vary throughout the range. For example, the seasonally varying photoperiod and climate means that populations need to be able to

predict the onset of unfavourable conditions (winter) from a changing photoperiod. However, the speed of the change in photoperiod is not the same at low and high latitudes imposing strong selection for local adaptation. The distribution of this species makes it an ideal system for the study of cold-tolerance in insects, diapause and the processes of adaptation to different environments.

As with many other *Drosophila* the male courtship song of *D. montana* is very important for mating success (Aspi & Hoikkala 1995), but also species recognition (Saarikettu et al., 2005). Populations differ in characteristics of the courtship song produced by males and in the preferences of females for particular characteristics (Ritchie et al., 1998; Klappert et al., 2007; Routtu et al., 2007; Ritchie et al., 2013). There is some evidence that these traits contribute to some amount of reproductive isolation between populations (Jennings et al., 2011). These traits are also highly dependent on the environment, in particular temperature, some of which may indicate the condition of males (Hoikkala et al., 2005). Some QTL are known for the variation in courtship song characters and female preference (Schäfer et al., 2010; Lagisz et al., 2012).

Populations also differ in important aspects of ecological adaptations to the environmental differences across latitude such as diapause behaviour (Tyukmaeva et al., 2011). In Finnish populations of *D. montana* spanning a latitudinal cline of 760km (~6˚ of latitude) critical day length (CDL), the length of the day (photoperiod) at which 50% of females enter diapause, is significantly correlated with latitude (Tyukmaeva et al., 2011). Another study in 2013 found the same trend but also showed that there was no apparent cline in the steepness of the photperiod response curve, the rate at which females in the population begin diapausing (Lankinen et al., 2013). Importantly, the CDL is the same after several generations of maintenance in the lab, indicating a genetic component to this phenotype (Lankinen et al., 2013).

D. montana is also known to be especially cold-tolerant among *Drosophila* species (Vesala & Hoikkala 2011; Vesala et al., 2012a; 2012b; 2012c). This cold tolerance is highly seasonal and related to the photoperiod, in the same way as diapause (Vesala et al., 2011; Vesala et al., 2012a). In lines from a high latitude Finnish population, shorter day lengths, as experienced naturally later in the summer an in fall, induce a period of cold acclimation (with shorter chill coma recovery times) and

diapause (Vesala, et al., 2012a). This pattern was not seen from a more southerly Canadian population indicating a lack of cold acclimation ability and cold-tolerance (though this strain had been in the lab for several generations) (Vesala et al., 2012a). Additionally, in some populations cold tolerance is enhanced when flies are in diapause, while in others it does not seem to have an effect (Vesala et al., 2011). In sum, *D. montana* shows some variation in cold tolerance among populations which is in some cases modulated by photoperiod induced diapause. Understanding how this variation is linked to the variation in climate and other environmental variables among populations will shed light on the process of population divergence and local adaptation in this species.

A major goal in evolutionary biology is to understand the genomics of population divergence and speciation (Butlin et al., 2012). Some studies have been done to investigate the molecular basis of these ecologically important traits in *D. montana*. Early studies probed the transcriptional profile of ~100-200 candidate genes on microarrays (Kankare et al., 2010; Vesala et al., 2012). These studies identified a number of genes which show changes in expression during cold acclimatisation and chill coma recovery. For example, expression differences between diapausing and non-diapausing females were observed for 24 out of 101 candidate genes, including the gene *Drosophila cold acclimation* (*Dca*) and *couch potato* (*cpo*) both of which have roles in cold acclimation and diapause in *D. melanogaster* respectively (Kankare et al., 2010). Another study found significant changes in many genes in response to cold hardening, cold acclimation and during chill coma recovery. Genes showing consistent changes include *period* (*per*) and various heat shock protein genes (Vesala et al., 2012). More recently, extensive screening by RNA-seq has corroborated these findings and uncovered further differences (Parker et al., 2015a; 2015b; Vigoder et al., 2016). In particular, the *myo-inositol-1-phosphate synthase* (*Inos*) gene shows upregulation in response to cold acclimation and is also the major metabolite present in overwintering flies (Parker et al., 2015a, Vesala et al., 2012b). Knockdown of this gene results in increased cold-induced mortality among flies (Vigoder et al., 2016). Additionally, some QTL have also been identified for cold tolerance and diapause (Tyukmaeva et al., 2015). Finally, a recent effort to produce more genomic resource to address these questions has produced a draft genome and linkage map for *D. montana* (Parker et al.,

in prep).

In summary, *D. montana* is characterised by a wide circumpolar distribution which extends into high latitudes, and high altitudes in the southern part of the range, which imposes seasonal and climatic selective pressures. Populations have diverged in their adaptations to these environmental conditions and also show some amount of reproductive isolation mediated by differences in courtship song and female preference functions (as well as some post-mating isolation). This system therefore represents an excellent model in which to study population divergence in the face of gene flow and the early stages of speciation. In this study I aim to assess the level of genomic differentiation among populations of *D. montana*. Given the species' polar distribution and cold-tolerance I also ask which genomic loci show differentiation that is also associated with clinal variation in climatic variables. Such SNPs should give insights into the extent of local adaptation among populations of *D. montana,* as well as alleles that contribute to barriers to gene flow (Butlin et al., 2012). I take advantage of recent genomic resources in *D. montana* and use a pool-seq approach to sample populations from parts of the geographic and climatic range of the species. The main aim of this work was to identify markers, and the nearby genes, where allele frequencies covaried with the environmental variation experienced by the species, including latitudinal and altitudinal variation in climate. Therefore, sampling was done at widely distributed populations which provided a range of latitude, altitude, and climate.

5.2 Materials and Methods

Samples of 49-50 individuals were collected in the spring of 2013 or 2014 from 4 populations along a latitudinal cline in North America (N.A.) and 2 populations along a latitudinal cline in Finland (figure 5.1; table 5.1). Populations cover a range of latitudes from 66 N to 38 N but one sample consitutes an outlier in terms of altitude. Crested Butte lies at an altitude of ~3,000m (table 5.1). Samples were stored in Ethanol prior to DNA extraction. DNA was extracted from individual flies using CTAB solution and phenol-chloroform-isoamylalcohol purifications. DNA concentrations were then measured with Qubit (Thermo Fisher Scientific) so that an equal amount of DNA from each individual (50 ng) was represented in the pooled sample. Sequencing was

performed at the Finnish Functional Genomics Centre in Turku, Finland (www.btk.fi/functional-genomics) on the Illumina HiSeq3000 platform (read length = 150bp, estimated coverage = ~121x).

Table 5.1. Characteristics of the populations sampled for pooled sequencing in this study. Data included are: the country and state/region (Source) of sampling, the name of the nearest town, coordinates and altitude of the sampling site, the year in which sampling was performed, the number of males and females sampled (M/F) from each site, and additional notes.

Source	***Sampling Site***	***Year***	***M/F***
Canada, Alaska	Seward 60º9'N; 149º27'W Altitude 35 m	2013	30/20
Canada, British Columbia	Terrace 54°27'N; 128°34'W Altitude 217 m	2014	22/27
USA, Washington	Ashford, 46°45'N; 121°57'W Altitude 573 m	2013	16/34
USA, Colorado	Crested Butte 38°54'N; 106°57'W Altitude 2900 m	2013	36/13
Finland	Oulanka 66°40′N; 29°20′E Altitude 337	2013	25/25
Finland	Korpilahti 62°20′N; 25°70′E Altitude 133	2013	27/23

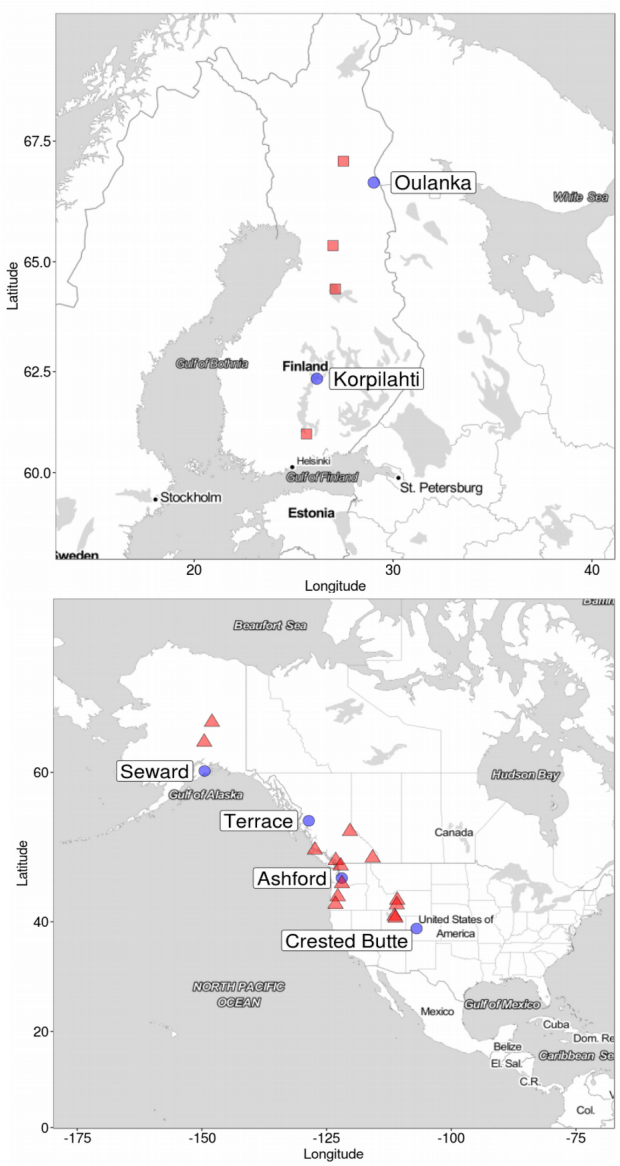


Figure 5.1. *D. montana* populations across Finland (Top) and North America (Bottom). Populations sampled for pooled sequencing are given in blue circles and named. Other populations in the clines are given by red squares (Top; Finland) or triangles (Bottom; North America). See table 5.1 for details. Maps are drawn using the "ggmap" (v. 2.7; Kahle & Wickham 2013) package in R.

5.2.1 Mapping and SNP Calling

Quality of reads was checked with FASTQC (v. 0.11.5) (Andrews 2014)**.** Reads

were trimmed using trimmomatic (v. 0.32) (Bolger *et al.* 2014)**.** A first round of trimming removed TruSeq3 adapters from the reads. Leading and trailing bases were removed if the quality of the base was below 20 and reads were cut if in a sliding window of 5 bases the quality fell below 20. Finally all reads < 100 bases long are discarded (Trim 1). Quality checks and the first round of trimming revealed failure of the "per base sequence content" modules indicating that the difference in allele frequencies at some positions were above 20%. This seemed to be driven by a drop-off of "A" bases at the ends of reads. Thus another set of trimming options were used which, in addition to all the steps in Trim 1, cropped reads at 145 bp (Trim2). This resulted in no failed reports from FASTQC and no drop off in the number of reads kept. The proportion of reads that made it through all trimming steps ranged between ~74 and ~81% of the original sample (table 5.2). Trimmed reads after Trim 2 were mapped to the *D. montana* reference genome (Parker et al., *in prep*) using BWA mem (v. 0.7.7) (Li 2013) with the default options but keeping only alignments with a mapping quality > 20 as per the best practice guidelines for pool-seq (Schlötterer et al., 2014). Duplicate alignments were removed with samtools rmdup (v 1.3.1) (Li et al., 2009). Regions around indels are re-aligned using picard (v. 1.118, Broad Institute *no date*) and GATK (v. 3.2-2, McKenna et al., 2010) and samtools. Separate .bam files for each sequencing sample were merged using bamtools (v. 2.4.0; Barnett *no date*).

Overall, mapping rates are good with over 80% of reads being properly mapped in all samples.Empirical coverage seems to be slightly lower (~100-110x) than the expected coverage (~120x). Coverage distributions for each sample are shown in figure 5.2. The mean coverage for the Seward samples was nearly twice that of the other samples (figure 5.2; table 5.3). After subsampling of Seward samples the distributions of coverage were much more similar among the populations, this allows common maximum and minimum coverage thresholds to be set based on the aggregate distribution of coverage (figure 5.2).

Sample coverage in terms of the number of reads, and reads per base, was nearly twice the coverage for other samples in Seward samples (table 5.2). To avoid the potential for this extreme difference in coverage causing artefacts in downstream analyses the .bam files for Seward were downsampled to contain 94.1 million reads which was the average across the remaining populations (when all samples from the

populations were combined; table 5.3, figure 5.2). Finally, allele frequencies among the pools were called with samtools mpileup (v. 1.3.1) using the options to skip indel calling as well as ignoring reads with a mapping quality < 20 and sites with a base quality < 15 (Li et al., 2009), the pileup file was converted to the .sync format using PoPoolation2 (v.1.2; Kofler et al., 2011).

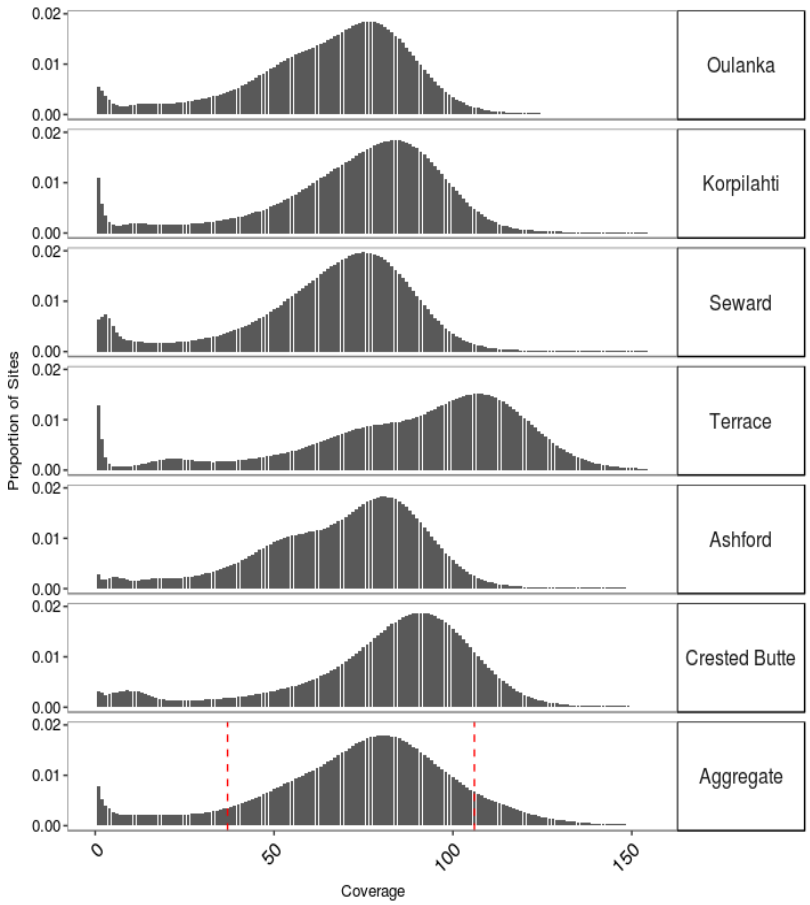


Figure 5.2. Coverage distributions for merged sequencing data from each of the populations as well as the "aggregate" distribution. Red vertical dashed lines give the 10th and 90th percentiles of the aggregate distributions used for setting coverage thresholds.

To avoid spurious results due to the physical linkage of SNPs we perform analyses only for scaffolds greater than 10kb in length. This excludes 60,317 scaffolds but keeps 76% of the total length of the genome. Furthermore, recent development of a linkage map resolves the ordering (but not orientation) of scaffolds across four major

linkage groups as well as the placement of some scaffolds on the X chromosome (Parker et al., *in prep*). Because an unequal number of males and females are used in most of the pools and coverage varies greatly across scaffolds and pools, the accurate estimation of allele frequencies on the X is difficult. For this reason, any scaffolds that could be assigned to the X chromosome linkage group are excluded from downstream analyses. We also consider only SNPs that can be reliably mapped to one of the other autosomal linkage groups. The final set contains 802,221 SNPs distributed across 835 scaffolds for further analysis. Linkage groups from this linkage map correspond well to the chromosomes in *D. montana* and *D. virilis* thus linkage groups are referred to as chromosomes throughout the text.

5.2.2 Climate Data

Representative climate data for each population sampled for pool-seq (see above) as well as 18 additional populations is obtained from the WorldClim database (Hijmans et al., 2005) using the longitude and latitude coordinates. The "bio" (Bioclimatic variables), "tmin" (Minimum temperature), "tmax" (Maximum temperature) and "prec" (Precipitation) data from the "Current conditions (~1960-1990)" dataset are obtained using the "raster" package (v. 2.5-8; Hijmans et al., 2016) in R (v. 3.3.2; R Development Core Team 2016) . In total this amounts to 55 bioclimatic variables for each population. To reduce the number of variables in the dataset a principle components analysis (PCA) is performed using the "PCA()" function from the "FactoMineR" package (v. 1.28; Lê et al., 2008) in R. Principle components are kept for further analysis if their eigenvalues are > 1. PCA scores for each population are z-transformed using the "scale()" function in base R. To determine whether the populations sampled for pool-seq are representative of the full range we repeated the PCA analysis using only these 6 populations.

Table 5.2. Number of reads before and after two difference trimming tests with parameter sets Trim 1 and Trim 2 (see main text for details). Read numbers are given for one direction only (R1), the total number of reads is twice the read count.

Sample	***Nr. Raw Reads***	***Trim 1: Nr. Reads Kept (% kept)***	***Trim 2: Nr. Reads Kept (% kept)***
Oulanka_S16_L002	43,716,441	33,659,994 (77.00)	33,659,994 (77.00)
Oulanka_S16_L003	40,627,318	30,048,541 (73.96)	30,048,541 (73.96)
Korpilahti_S11_L002	46,443,624	37,031,413 (79.73)	37,031,413 (79.73)
Korpilahti_S11_L003	43,070,443	37,031,413 (85.98)	33,665,807 (78.16)
Ashford_S12_L002	41,165,726	32,756,890 (79.57)	32,756,890 (79.57)
Ashford_S12_L003	37,393,574	29,159,469 (77.98)	29,159,469 (77.98)
Crested_Butte_S13_L002	50,741,697	40,986,231 (80.77)	40,986,231 (80.77)
Crested_Butte_S13_L003	51,249,585	40,604,980 (79.23)	40,604,980 (79.23)
Seward_S14_L002	100,675,666	81,842,768 (81.29)	81,842,768 (81.29)
Seward_S14_L003	102,610,964	81,702,658 (79.62)	81,702,658 (79.62)
Terrace_S15_L002	56,340,618	45,742,952 (81.19)	45,742,952 (81.19)
Terrace_S15_L003	59,591,845	47,456,742 (79.64)	47,456,742 (79.64)

Table 5.3. Post-mapping statistics from samtools flagstat and bedtools coverageBam after merging of .bam files and subsampling of Seward samples.

Sample	***Nr. Properly Paired Reads (% of total)***	***Mean Coverage***
Oulanka	78,576,411 (86.0)	90.01x
Korpilahti	88,619,130 (86.4)	99.40x
Ashford	82,605,252 (86.7)	92.96x
Crested Butte	97,825,698 (86.9)	107.68x
Seward	79,962,199 (87.27)	88.59x
Terrace	109,657,304 (86.8)	118.14x

5.2.3 Clinal Analysis

Several possible approaches to the analysis of the relationship between allele frequencies and environmental gradients exist. Here we apply three recent methods that control for various sources of error and population history.

Quasibinomial GLMs

For each SNP the allele frequency of the "A" allele within each population is estimated. Allele frequencies were used as the response in a generalised linear model (GLM) with a quasibinomial error distribution (see Chapter 2). Principle components from the PCA were used as predictor variables. The model takes the form:

$$y = altitude + PC1 + PC2 + e$$

Where "*y"* are the allele frequencies in each population given as counts of the major and minor alleles, the predictors are the climatic variables from a PC analysis (PC1 and PC2), as well as the altitude of each population, and "*e*" is a quasibinomially distributed error term. In this analysis allele frequencies were scaled to the effective sample size (n_{eff}; Kolaczkowski et al., 2011; Feder et al., 2012). First the frequency of the major allele (f_A) in the pool was determined. Then n_{eff} was calculated as:

$$n_{eff} = (cov * n - 1)/(cov + n),$$

where *cov* is the total coverage at the SNP, *n* is the number of chromosomes in the pool ($n = 2N = 2*50 = 100$). Then new counts for the major (A_c) and minor (a_c) allele were computed from the new effective sample size by:

$$Ac = fA * neff$$

and;

$$ac = neff - Ac$$

rounding off to the nearest integer.

The quasibinomial GLM tests for a linear relationship between the allele

frequencies at each SNP and the continuous variation in climatic predictor variable regardless of continent. The model is controlling for a partial effect of altitude, however the effect of altitude is probably unreliable because the Crested Butte population is a strong outlier (table 5.1).

In the above models only alleles with a minimum count of 10 reads (across all populations) are considered. Additionally, a minimum and maximum coverage of 37 and 107 (in all populations) are imposed. Quasibinomial GLMs were fitted to the allele frequencies for each SNP using custom python and R scripts (available from: https://github.com/RAWWiberg/ThCh5).

BayeScEnv

Recent developments in the analytical methods to identify outlier loci based on F_{ST} have produced several software tools (Foll & Gaggiotti 2008; de Villemereuil & Gaggiotti 2015). In particular, the need to identify loci that underlie local adaptation in various contexts motivated the software BayeScEnv (de Villemereuil & Gaggiotti 2015). BayeScEnv uses a population genetic model of population differentiation and tests for a relationship between genetic differentiation (F_{ST}) and environmental differentiation across populations (de Villemereuil & Gaggiotti 2015). The method compares models with and without the additional term giving a relationship between environmental and genetic differentiation. This way the signals of differentiation above that expected from demographic effects (differences in N_e or migration) or a model of other locus-specific effects (e.g. background selection, variation in mutation rates; de Villemereuil & Gaggiotti 2015) are accounted for. In this study, the the first two PCs (see above) were entered one at a time as the environmental explanatory variable. The input for BayeScEnv are the allele frequencies for each SNP, given as counts of the major and minor alleles, within all 6 populations as well as the environmental variables from each population. As with the quasibinomial GLMs (see above), the allele counts were scaled to n_{eff}. Input files for BayeScEnv were generated using in-house scripts (available from: https://github.com/RAWWiberg/ThCh5).

5.2.4 Functional Genomic Analysis

The closest genes to the SNPs showing a significant relationship to

environmental differenatiation among the populations was extracted from the *D. montana* annotation with the closestBed routine from bedtools (v. 2.17.0; Quinlan & Hall 2010). The *D. virilis* orthologs of these *D. montana* genes were submitted to DAVID (v. 6.8; Huang et al., 2009a; 2009b). A functional cluster was considered significant if the enrichment score (ES; the geometric mean of -log(p-values) of GO terms in the functional group) was > 1.3, corresponding to a mean p-value < 0.05 (Huang et al., 2009b).

The gene lists were also submitted to the phenotype enrichment analysis software DroPHEA (Weng & Liao 2011) which tests for an enrichment of genes in different mutant phenotype categories. The gene lists were also manually checked for genes previously associated with latitudinal clines or relevant phenotypes known to vary clinally in *D. montana*.

Finally, I take advantage of data from recent RNA-seq experiments which have uncovered a number of genes showing differential expression (DE) under cold acclimation (Parker et al., 2015), diapause (Kankare et al., 2016), and changes in the photoperiod (Parker et al., 2016). I ask how much overlap there is between DE genes in these studies and those genes closest to SNPs which show significant association with environmental variables. Using a simple bootstrap approach to ask what amount of overlap is expected by chance if an equal number of genes are drawn from the *D. montana* annotation at random. Only genes that for which a reliable ortholog in *D. virilis* is identified are used from these published datasets.

5.3 Results

5.3.1 Mapping and SNP Calling

The total number of SNPs in the dataset was 2,980,157. From the data available on linkage groups (Parker et al., *in prep*) some scaffolds can be assigned to chromsomes in their appropriate order (but not orientation). After removing small (< 10kb) scaffolds, 1,195 scaffolds (with a mean length of ~40kb can be arranged onto four autosomal chromosomes and on one X-chromosome linkage group (herafter simply referred to as chromosomes). Due to the difficulty of accurately estimating allele frequencies from pools of individuals where different numbers of males and females are sequenced the

scaffolds mapped to the X-chromosome are excluded from the below analyses. The total number of SNPs on scaffolds > 10kb and not on X-linked scaffolds was 2,559,863 SNPs. Of these, 802,221 SNPs are distributed across the scaffolds arranged on four chromosomes. Analyses and multiple test correction were performed for the full set of 2,559,863 SNPs.

5.3.2 Climate Data

PC analysis (PCA) of the WorldClim climate data was performed for all 24 populations from which *D. montana* have been identified and climate data were collected. These results revealed four principle components (PCs) that together exaplain ~98% of the variation (figure 5.3). The first two principle components seem to split the populations roughly first by a measure of "distance inland" (PC1) and then by latitude (or altitude) (PC2). PC1 explains ~55% of the variation (figure 5.3 and figure 5.4) and loads heavily on climate and biological variables that are associated with precipitation and temperature e.g. "Mean Temperature of Coldest Quarter", "Precipitation of Wettest Month", "Annual Precipitation". This maps intuitively on to the populations. The highest scoring population for PC1 is Ashford (figure 5.4) which is a southern population on the pacific coast and as such receives most rain but also experiences warm summers and mild winters. Meanwhile, PC2 explains ~23% of the variation and loads heavily on biological variables that are associated with latitudinal clinality, e.g. "Mean Diurnal [Temperature] Range," and "Isothermality" which is the diurnal range divided by the mean "Annual [Temperature] Range." This also maps onto populations quite intuitively; high scoring populations have higher latitude (figure 5.1, figure 5.4, table 5.1). The remaining principle components (PC3 and PC4) explain ~11.5 and 5% of the variation respectively and are not capturing as much of the climatic variaiton as the first two components.

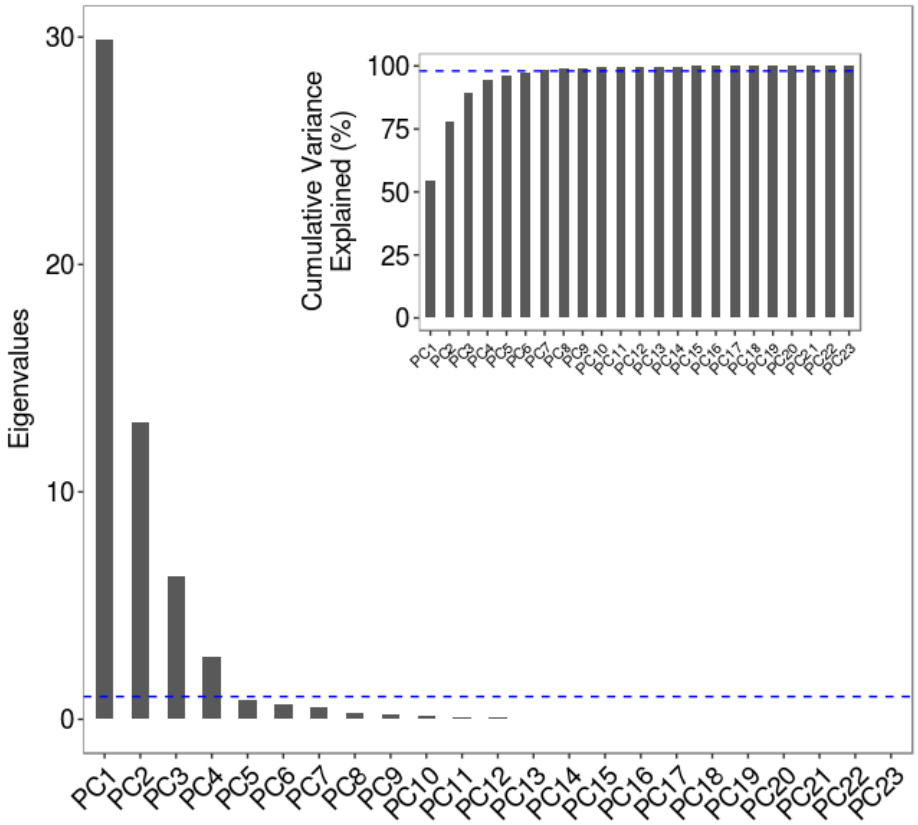


Figure 5.3. Barplot showing the eigenvalues of the principal components (PCs) The dashed blue line in the main figure delineates the threshold for including a PC in further analyses and represents and eigenvalue of 1. The inset figure shows the cumulative variance explained by each additional principal component. The blue dashed line in the inset figure delineates 98% of the variation explained. The data shown are for a PC analysis with all 24 *D. montana* populations for which climate data were collected.

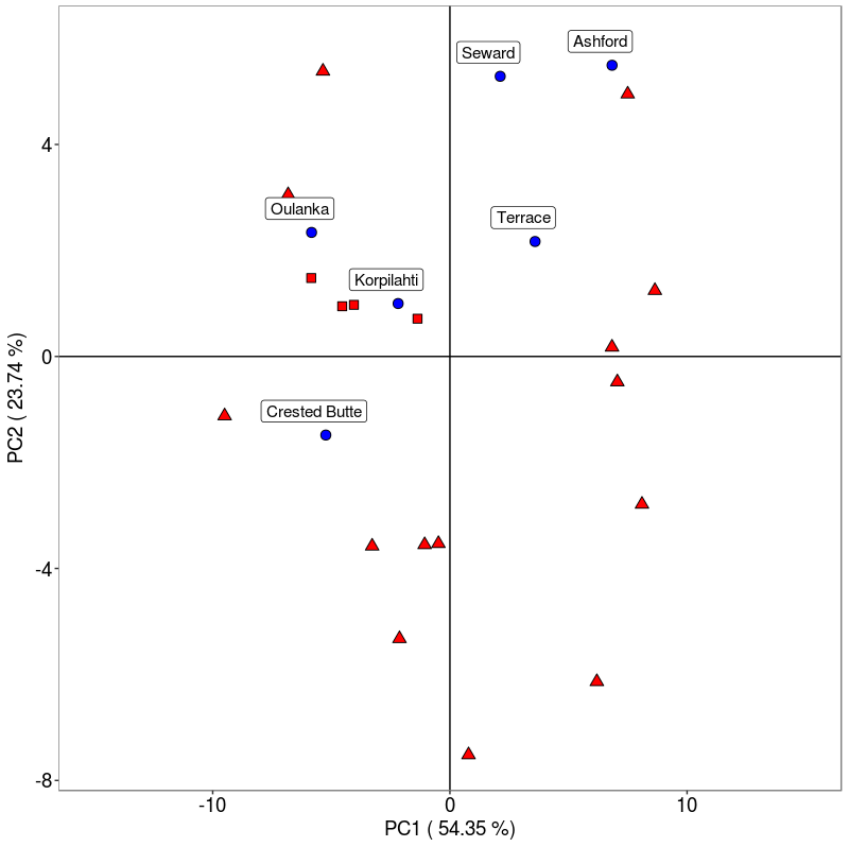


Figure 5.4. 1st and 2nd principle components (PC1 and PC2) from a PCA analysis of 54 climate variables from WorldClim (Hijmans et al., 2005). PC1 explains 54.35% of the variance while PC2 explains 23.74%. Populations sampled for pool-seq are given as blue circles and named while the remaining populations are given as red squares (Finnish populations) or triangles (North American populations).

If the data are subset to only the 6 populations sampled for pool-seq the first four PCs still explain ~98% of the variation**.** In fact, 29 of 32 variables that load strongly on PC 1 also load strongly on PC 1 in the first analysis including all populations. Similar results hold for PC 2 where two out of 5 strongly loading variables also load strongly in the first analysis. The two first PCs also explain similar amounts of the variation in both analyses. PCs 1 and 2 explain ~59 and 21% of the variation respectively in the latter analysis (figure 5.5). PCs 1 and 2 also map intuitively to the geographical locations of the populations. A higher score on PC1 is again associated more inland populations while a lower score on PC2 is associated with higher latitudes (or altitude in the case of

Crested Butte). As before, the remaining PCs explain much less of the variation and therefore do not capture much of the environmental variation. Taken together, these results suggest the subset of populations that were sampled for pool-seq give a very good representation of the climatic variation experienced by this species throughout it's range. Thus, any relationship between environmental variables and genetic differentiation in the samples selected for pool-seq is more likely to reflect true patterns across the populations. For all downstream analyses the PC1 and PC2 values, from the PCA including all populations, for each of the populations sampled for pool-seq were used as explanatory variables in the quasibinomial GLMs or as input to BayScEnv (see below).

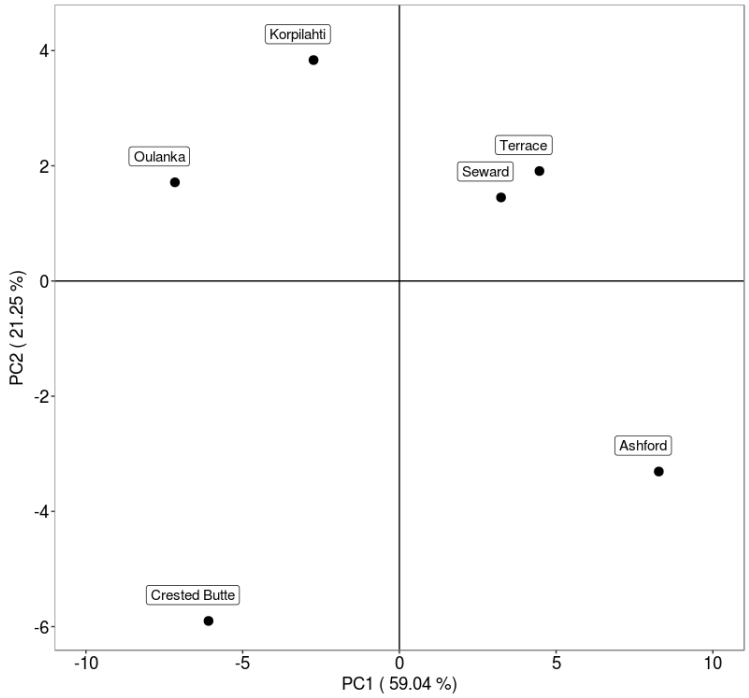


Figure 5.5. Principle components (PCs) 1 and 2 from a PCA of 55 bio-climatic variables for the 6 populations that were sampled for pool-seq.

5.3.3. Patterns of Genetic Diversity

Figure 5.6 gives the overall levels of nucleotide diversity (π) and Tajima's D for each population in this study. Generally diversity is a little bit lower among the two

norther Finnish populations than among the North American populations, perhaps indicating a relatively recent bottleneck. One outlier is the Crested Butte population which has substantially lower diversity than other populations. These patterns are seen also in Tajima's D (figure 5.6) which show a strongly negative value for Crested Butte indicating an excess of rare alleles genome-wide.

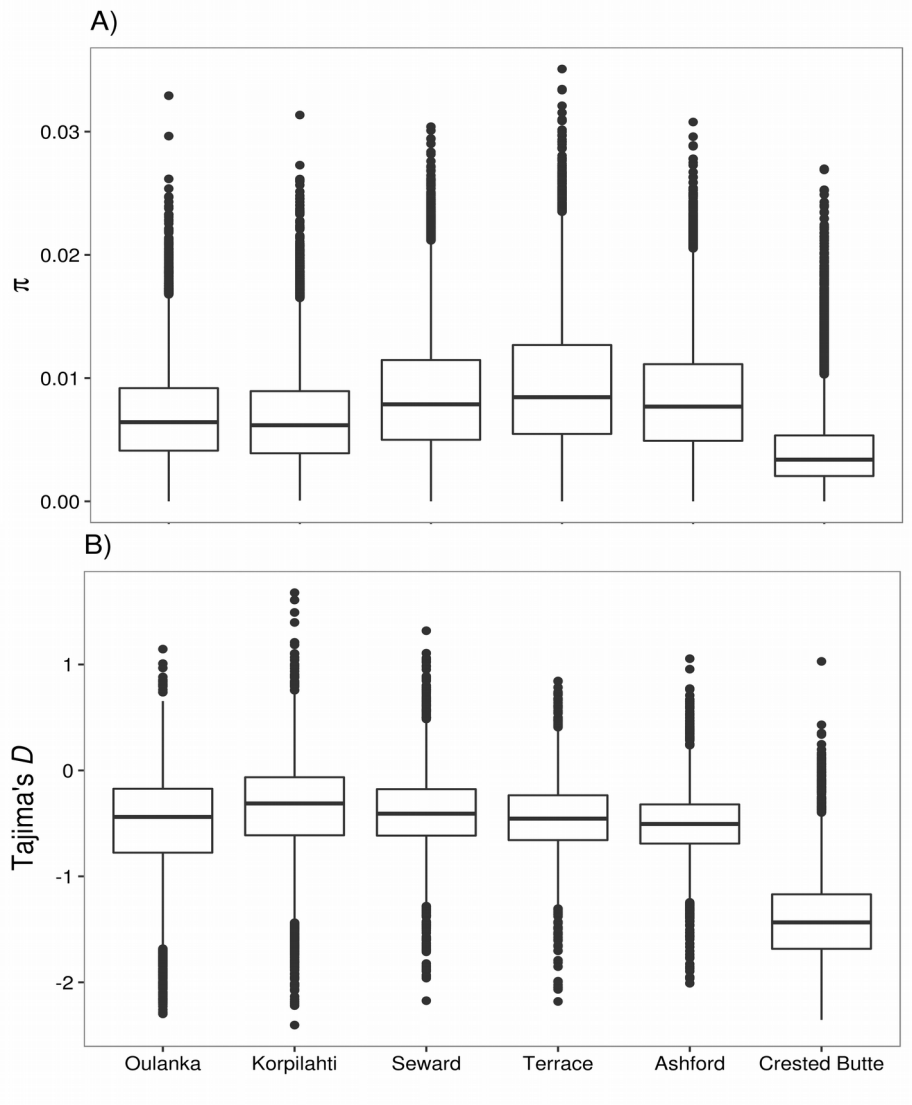


Figure 5.6. Estimates of A) nucleotie diversity (π) and B) Tajimas'D. Calculations were made within non-overlapping 10kb windows. Calculations were performed in PoPoolation (v. 1.2; Kofler et al., 2011)

5.3.4 Clinal Analysis:

Quasibinomial GLMs

Results from quasibinomial GLMs suggest that they perform rather poorly in this study. P-value distributions are in some cases substantially inflated for intermediate p-values (figure 5.7). Such a distribution does not allow for multiple test correction by conventional methods (e.g. q-values) and usually reflect a poor fit of the model to the data. For this reason we consider them unreliable (see discussion for potential explanations). Indeed, attempting to correct for multiple testing by q-values results in no significant hits (q-values < 0.05) (figure 5.7 D).

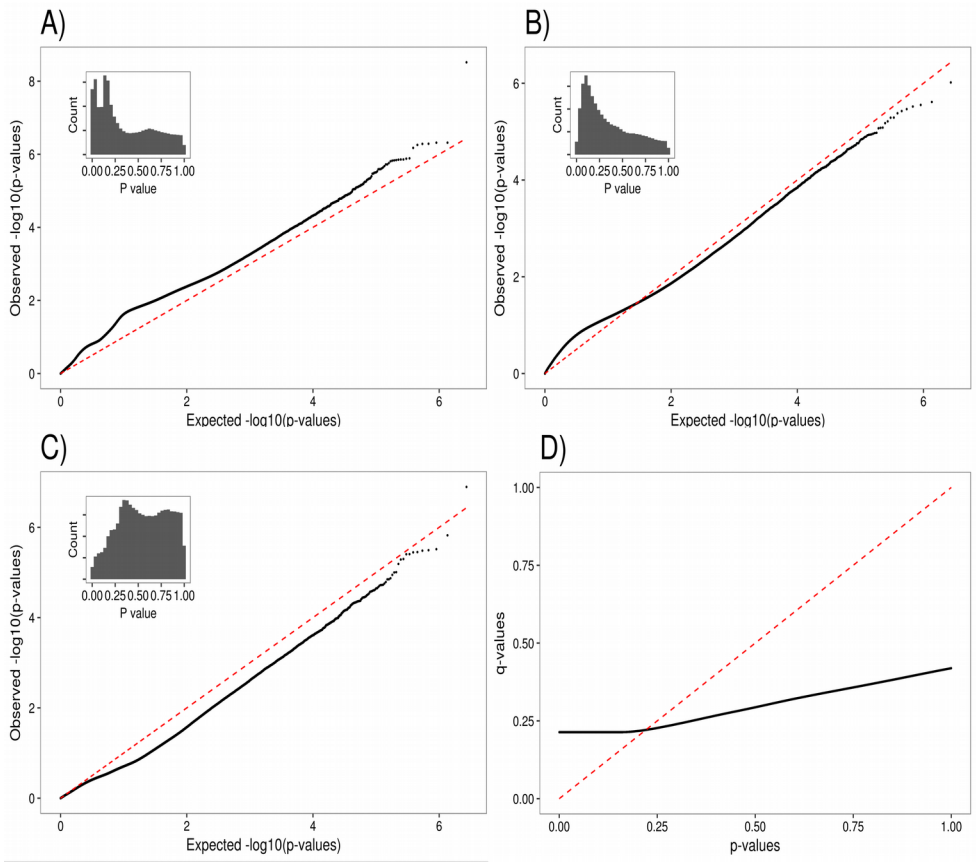


Figure 5.7. log-log plots of expected and observed p-values from quasibinomial GLMs with **A)** altitude, **B)** PC1 and **C)** PC2 as explanatory variables. Inset figures in **A)**, **B)** and **C)** show the distribution of p-values. **D)** gives a summary of the p-value to q-value conversion for PC1 as an illustration.

BayeScEnv

In total, 2,559,863 SNPs are analysed in BayeScEnv. Results from BayeScEnv find several SNPs with a significant relationship between F_{ST} and environmental differentiation among populations (figures 5.8 and 5.9). For PC1 4,095 SNPs are significantly associated with the environmental variable (figure 5.8). The top SNPs are not evenly distributed among the chromosomes. The total number of SNPs is uneven across the major autosomal chromosomes (χ^2=23880, d.f. = 3, $p < 0.001$) and this trend does not seem to be related to the length of the chromosomes, i.e. there is no clear relationship between the length of the chromosome and the number of SNPs (although there are too few data points for any meaningful correlation test). Despite this, the distribution of the number of significant SNPs is also highly skewed (χ^2=2016.7, d.f. = 3, $p < 0.001$). This pattern holds if the observed proportions of all SNPs on the linkage groups are used as the expected proportions of top SNPs (χ^2=1691.3, d.f. = 3, $p < 0.001$). While chromosome three has the most SNPs in total, chromosome four has far more significant SNPs (2,255) than the other chromosomes (Chr 2: 470 SNPs, Chr 3: 674, Chr 5: 606; figure 5.8). Several SNPs also show a significant relationship with PC2 (figure 5.9). Many of the peaks seem to be shared between PC1 and PC2 (figures 5.8 and 5.9). Indeed the set of significant SNPs overlap by 26.9% (N = 1,937). PC1 has 2,158 (30% of all SNPs) private SNPs and PC2 has 3,106 (43.1% of all SNPs) private SNPs.

5.3.5 Functional Analysis

The significant SNPs from BayeScEnv analysis of PC1 all lie within 1Mb of annotated genes. In total, there are 731 unique genes within 1Mb of the significant SNPs. Meanwhile, there are 713 genes within 1Mb of the significant SNPs in the analysis of PC2. The overlap in the genes for PC1 and PC2 is high (508; 54%). PC1 has a unique set of 223 genes and PC2 has a unique set of 205 genes. DAVID enrichment analysis of these sets indicates that they are strongly enriched for several annotation clusters (table 5.4) including ion transport, transmembrane proteins, lipoproteins and lipases (table 5.4). Meanwhile, these same genes show a significant enrichment for phenotypic classes such as “eclosion defective”, “phototaxis defective”, “neurophysiology defective”, and “hyperplasia.”

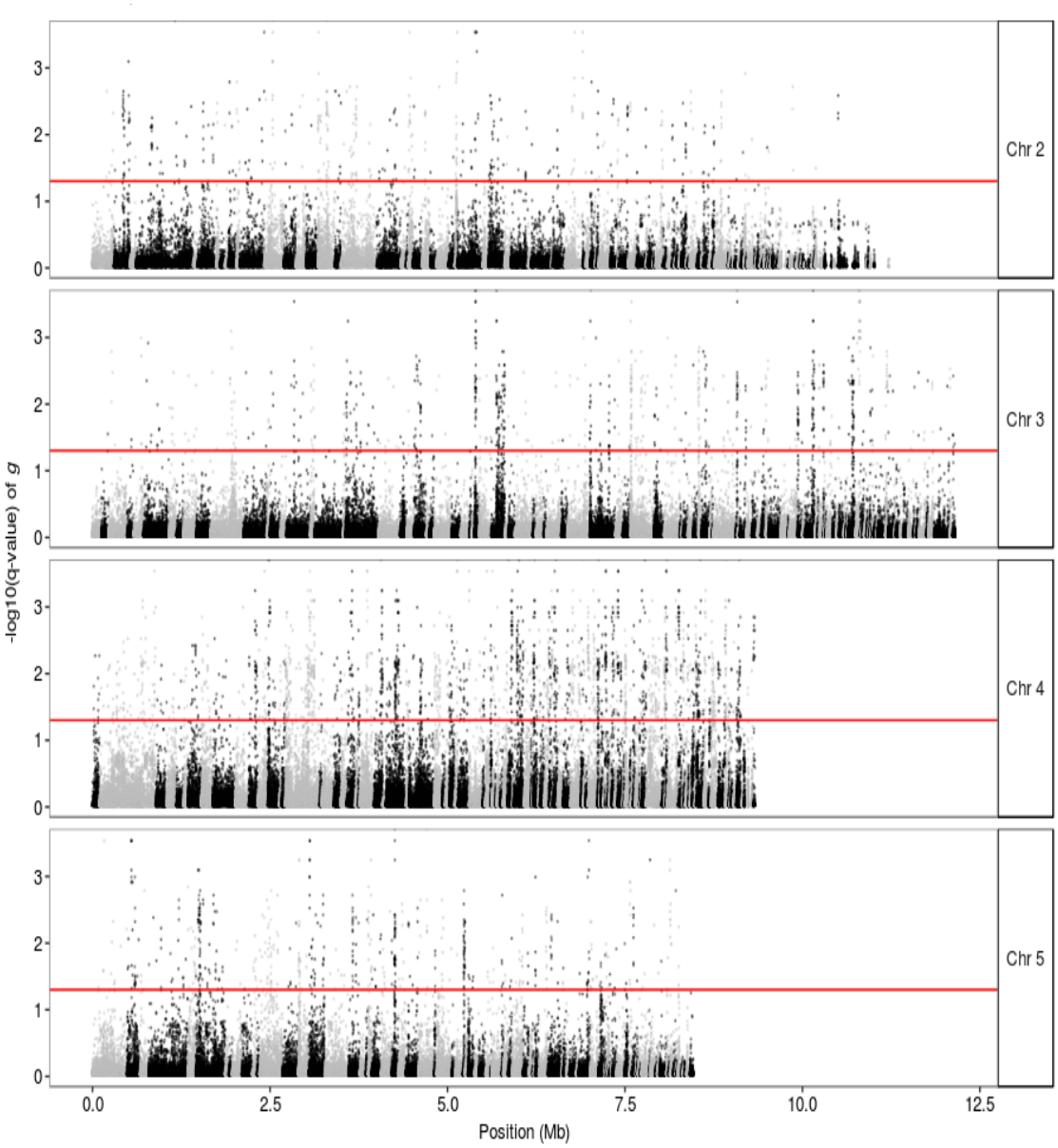


Figure 5.8. Manhattan plot of -$\log_{10}$(q-values) for *g* parameter in BayeScEnv which shows the degree of association between F_{ST} at a SNP and PC1. Alternating grey and black coloured points show different scaffolds. The red horizontal line shows the q-value = 0.05 threshold. Panels are the different linkage groups (chromosomes) to which scaffolds can be anchored.

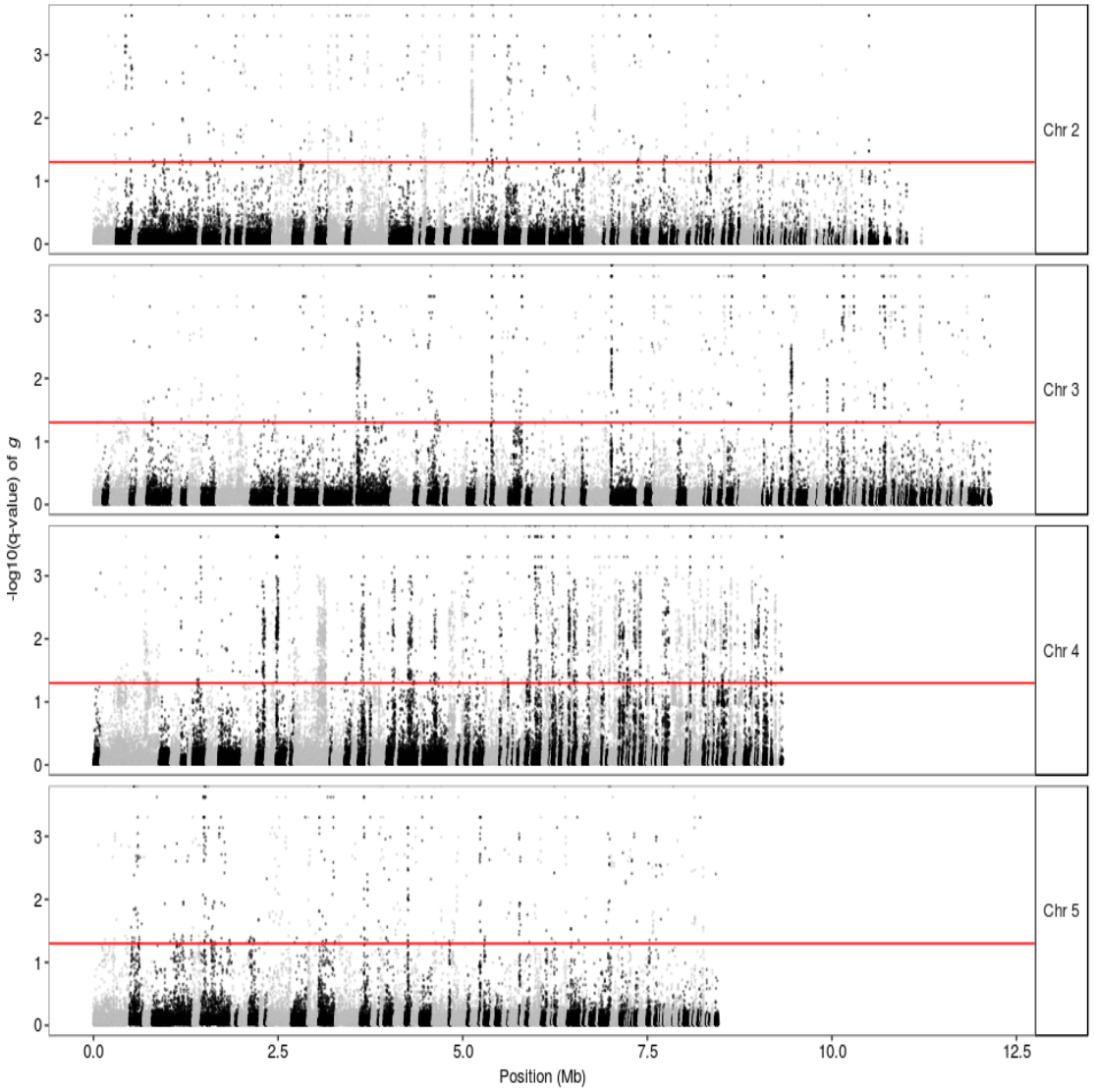


Figure 5.9. Manhattan plot of -$\log_{10}$(q-values) for the *g* parameter in BayeScEnv which shows the degree of association between F_{ST} at a SNP and PC2. Alternating grey and black coloured points show different scaffolds. The red horizontal line shows the q-value = 0.05 threshold. Panels are the different linkage groups (chromosomes) to which scaffolds can be anchored.

Table 5.4. Significant DAVID functional clusters of genes within 1Mb of SNPs significantly associated with PC1 and PC2 in BayeScEnv. Given are the GO terms within each cluster along with their individual Benjamini-Hochberg corrected p-values as well as the enrichment score (E). The table is split into enrichment analysis for the genes common to PC1 and PC2, those unique to PC1, and those unique to PC2. Each cluster contains several GO terms as defined by different sources, these sources are given in square brackets. E is the geometric mean of individual GO term p-values, in $-\log_{10}$ scale, these means are given in brackets with the enrichment score.

Cluster	***GO terms (Adj. p-value)***	***E***
Clusters from Genes Common to PC1 and PC2		
1	[INTERPRO] Immunoglobulin V-set (0.012) [INTERPRO] Immunoglobulin subtype 2 (0.14) [INTERPRO] Immunoglobulin subtype (0.14) [INTERPRO] CD80-like, immunoglobulin C2-set (0.13) [SMART] IGc2 (0.18) [SMART] IG (0.14) [INTERPRO] Immunoglobulin-like fold (0.53) [UP_KEYWORDS] Immunoglobulin domain (0.21) [INTERPRO] Immunoglobulin I-set (0.9)	2.86 (0.001)
2	[UP_KEYWORDS] Lipoprotein (0.16) [UP_KEYWORDS] Glycoprotein (0.20) [UP_KEYWORDS] GPI-anchor (0.22) [GOTERM_CC_DIRECT] anchored component of membrane (0.71)	1.72 (0.02)
3	[UP_KEYWORDS] Membrane (0.18) [UP_KEYWORDS] Transmembrane helix (0.21) [UP_KEYWORDS] Transmembrane (0.20) [GOTERM_CC_DIRECT] integral component of membrane (1.0)	1.51 (0.03)
4	[INTERPRO] Sodium (0.78) [GOTERM_MF_DIRECT] neurotransmitter: sodium symporter activity (0.98) [UP_KEYWORDS] Symport (0.42)	1.47 (0.03)

5	[UP_KEYWORDS] Potassium (0.27) [UP_KEYWORDS] Transport (0.18) [UP_KEYWORDS] Voltage-gated channel (0.18) [INTERPRO] Potassium channel, voltage dependent, Kv (0.67) [UP_KEYWORDS] Potassium channel (0.21) [UP_KEYWORDS] Ion channel (0.21) [GOTERM_CC_DIRECT] voltage-gated potassium channel complex (0.85) [UP_KEYWORDS] Ion transport (0.21) [INTERPRO] Voltage-dependent potassium channel, four helix bundle domain (0.95) [INTERPRO] Potassium channel tetramerisation-type BTB domain (0.97) [GOTERM_BP_DIRECT] protein homooligomerization (1.0) [UP_KEYWORDS] Potassium transport (0.34) [GOTERM_MF_DIRECT] voltage-gated potassium channel activity (1.0) [INTERPRO] Ion transport domain (0.99) [INTERPRO] BTB/POZ-like (0.99) [INTERPRO] BTB/POZ fold (0.99) [SMART] BTB (1.0)	1.42 (0.04)
Clusters from Genes Unique to PC1		
1	[INTERPRO] DnaJ domain (0.83) [INTERPRO] DnaJ domain, conserved site (0.64) [SMART] DnaJ (0.83)	2.09 (0.008)
2	[INTERPRO] DnaJ domain (0.83) [INTERPRO] Chaperone DnaJ, C-terminal (0.76) [INTERPRO] HSP40/DnaJ peptide-binding (0.76) [GOTERM_BP_DIRECT] protein folding (1.0)	1.7 (0.02)
3	[INTERPRO] Leucine-rich repeat (0.77) [INTERPRO] Lecine-rich repeat, typical subtype (0.95) [SMART] LRR TYP (0.80)	1.47 (0.03)

4	[SMART] EGF (0.74) [INTERPRO] Epidermal growth factor-like domain (0.95) [INTERPRO] EGF-like, conserved site (0.99)	1.31 (0.05)
5	[INTERPRO] Lipase (0.86) [INTERPRO] Lipase, N-terminal (0.86) [GOTERM_BP_DIRECT] lipid metabolic process (0.96) [GOTERM_BP_DIRECT] extracellular region (0.96) [UP_KEYWORDS] Secreted (0.88)	1.31 (0.05)
Clusters from Genes Unique to PC2		
1	[UP_KEYWORDS] Leucin-rich repeat (0.012) [INTERPRO] Toll/interleukin-1 receptor homology (TIR) domain (0.98) [SMART] TIR (0.70) [INTERPRO] Leucine-rich repeat (0.99) [INTERPRO] Leucine-rich repeat, typical subtype (0.93) [SMART] LRR TYP (0.91) [GO_TERM_DIRECT] signal transduction (1.0)	1.8 (0.02)
2	[INTERPRO] Ankyrin repeat (0.92) [INTERPRO] Ankyrin repeat containing domain (0.88) [SMART] ANK (0.82) [UP_KEYWORDS] ANK repeat (0.85)	1.31 (0.05)

Overlap of the genes identified here and those identified in previous studies is relatively low. Only 7 of the 936 unique genes (0.7%) are also reported as DE in response to cold acclimation from Parker et al., (2015). While this overlap is low, it is greater than expected by chance (empirical p-value < 0.01 from 1,000 bootstrap samples of 936 genes). The overlap with DE genes from a study on whole diapausing or non-diapausing flies showed more overlap (~66%), again this was much greater than expected by chance (empirical p-value < 0.01 from 1,000 bootstrap samples of 939 genes). Finally, comparison with a set of genes that are DE in response to changing light conditions in diapausing and non-diapausing flies (Parker et al., 2016) gives a lower

overlap of ~1%. However, this is still significantly greater than expected by chance (empirical p-value < 0.05 from 1,000 bootstrap samples of 939 genes). In total, there are 11 genes which occur in at least two of these transcriptome datasets (table 5.5). If these genes are consistently under selection in parts of the *D. montana* range then signatures of selective sweeps or background selection might be expected (e.g. negative Tajima's *D*). Figure 5.9 shows the patterns of Tajima's *D* for 100kb up- and down-stream of the focal genes in table 5.5. Despite the low resolution and the difficulty in determining the orientation of scaffolds on the linkage groups, in some cases there seems to be a reduction in Tajima's *D* in the most northern populations (Oulanka and Korpilahti, Finland) and the high altitude population (Crested Butte, Colorado) when compared to the other populations. For example, the gene *Sterile20-like kinase* (*Slik)* co-localises with a trough of Tajima's *D* in Oulanka, Korpilahti, and Crested Butte populations while Ashford, Terrace and Seward show values of Tajima's D closer to zero (figure 5.10). A second gene, *yolk protein 3* (*Yp3*), is also associated with such a pattern of Tajima's D (figure 5.10).

Several other genes also lie near SNPs with an association to PC1, PC2 or both. While they occur within 1Mb of SNPs showing an association with climatic variables some of the patterns of Tajima's *D* around the genes also show the characteristic footprint of selective sweeps or background selection. For example upstream of the gene *glass* (*gl*) is a strong dip in Tajima's *D* across all populations (figure 5.11). Similar pattern is seen near the *inactivation no afterpotential C* (*inaC*) and *sine oculis* (*so*) which both localise to a small region on the same scaffold (figure 5.11). However, *inaC* is linked to SNPs associated with PC1 while *so* is linked to SNPs assocated with PC2.

Table 5.5 The genes that are found near SNPs showing a relationship with PC1, PC2, or both which are also differentially expressed (DE) in at least two of the previous gene expression studies.

Gene	***Function***	***SNPs***	***Evidence of DE in:***
CG42313	Unknown	PC1 and PC2	Kankare et al., 2016; Parker et al., 2016
Pex12	Peroxisome protein; Sperm development	PC1 and PC2	Kankare et al., 2016; Parker et al., 2016
CG9008	Unknown	PC1 and PC2	Parker et al., 2015; Kankare et al., 2016
Yp3	Embryo development	PC1 and PC2	Kankare et al., 2016; Parker et al., 2016
Slik	Regulation of mitosis; Cell proliferation in imaginal disc.	PC1 and PC2	Kankare et al., 2016; Parker et al., 2016
Dvir\GJ18078	Unknown	PC1	Kankare et al., 2016; Parker et al., 2016
Inos	*Myo*-inositol synthesis; Cold tolerance	PC1	Parker et al., 2015; Kankare et al., 2016
Ltn1	Zinc ion binding; sleep	PC2	Kankare et al., 2016; Parker et al., 2016
Dvir\GJ16316	Unknown	PC2	Kankare et al., 2016; Parker et al., 2016
Vri	Transcription factor; Circadian rhythm; repressor of *Clk* and *cry*	PC2	Parker et al., 2015; Kankare et al., 2016
Sap47	Required for synaptic and behavioral plasticity	PC2	Kankare et al., 2016; Parker et al., 2016

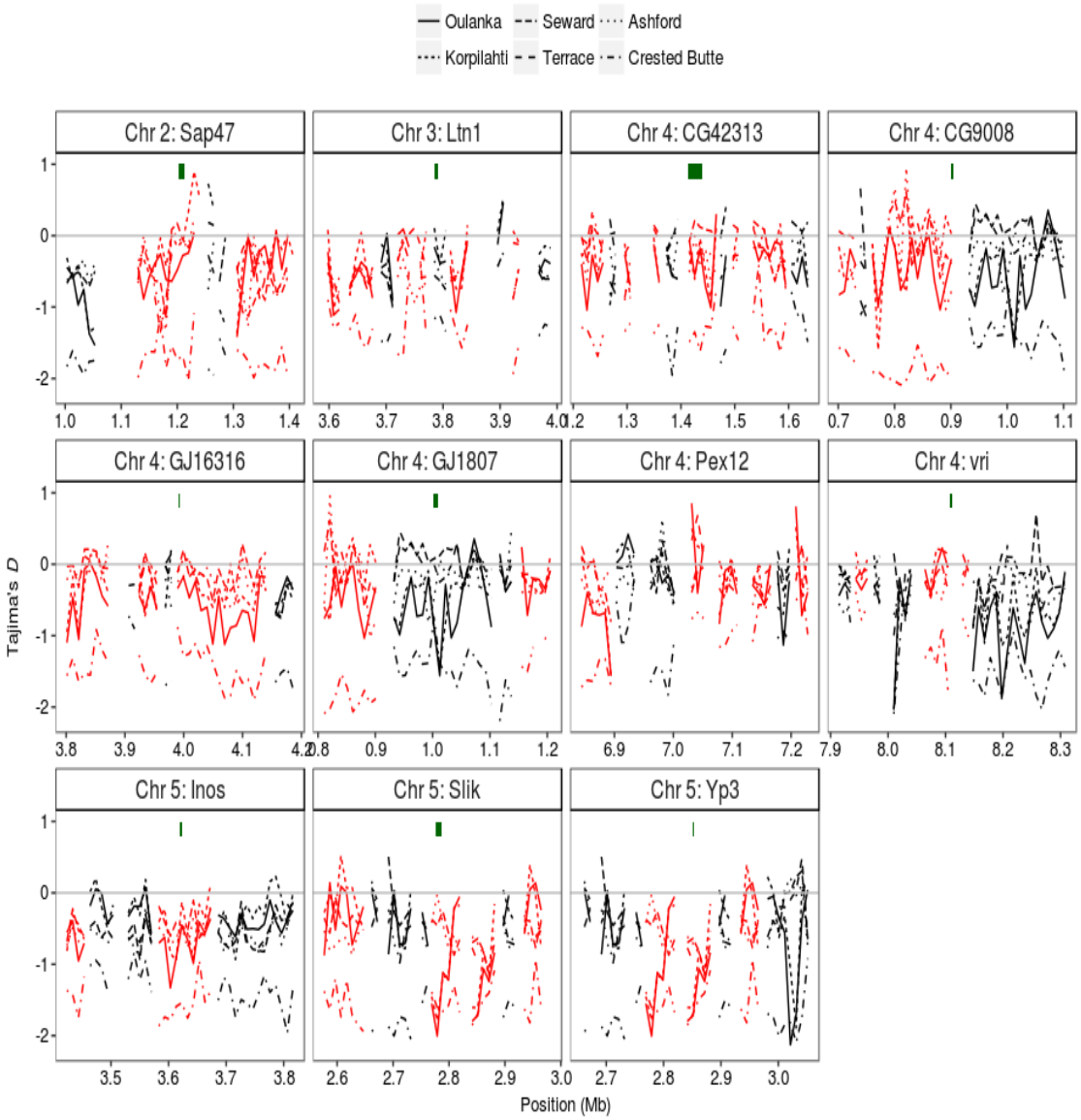


Figure 5.10. Patterns of Tajima's D around the focal genes in table 5.5. The genic region is given by a dark green bar at the top of each panel. Alternating colours of lines show different scaffolds. Scaffolds are placed in the correct order but the orientation of each scaffold is unknown. Different line types show the different populations. Tajima's D was calculated using PoPoolation2 (v. 1.2; Kofler et al., 2011)

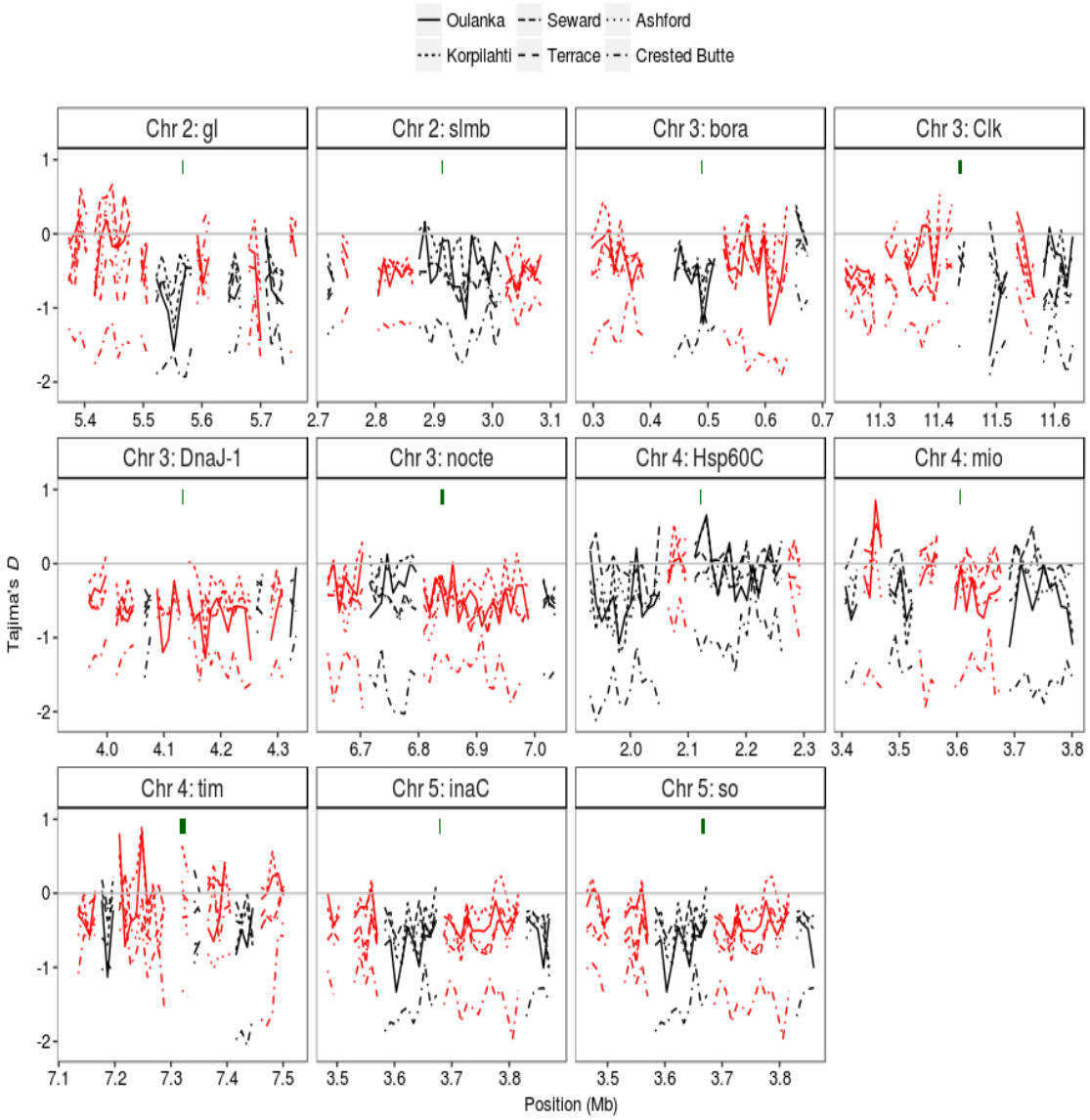


Figure 5.11. Patterns of Tajima's D near focal genes that occur near SNPs associated with the climatic variables PC1, PC2 or both. The genic region is given by a dark green bar at the top of each panel. Alternating colours of lines show different scaffolds. Scaffolds are placed in the correct order but the orientation of each scaffold is unknown. Different line types show the different populations. Tajima's D was calculated using PoPoolation2 (v. 1.2; Kofler et al., 2011)

5.4 Discussion

Identifying loci underlying local adaptation and population divergence will provide useful insights into the process of population differentiation and speciation. A fruitful approach is to study populations that vary continously in climatic or latitudinal clines (Endler 1973; Kolaczkowski et al., 2011; Cheng et al., 2012; Bergland et al., 2014; Machado et al., 2015; Takahashi 2015; Kapun et al., 2016). The fruitfly *D. montana* is distributed throughout the northern hemisphere and is found at higher altitudes further south (Throckmorton 1982). It is one of the most cold-tolerant species of fruitflies and overwinters in adult reproductive diapause (Throckmorton 1982). Previous population genetic studies have not found substantial genetic differentiation at microsatellite markers among populations of *D. montana* in Finland, indicating substantial gene flow (Tyukmaeva et al., 2011). At the same time there are differences between populations in various ecologically important traits like diapause and cold-tolerance (Tyukmaeva et al., 2011; Vesala & Hoikkala 2011; Vesala et al., 2012a; 2012b; 2012c). This variation among populations in important, adaptive traits coupled with the environmental variation throughout its range presents a valuable opportunity for understanding the forces that shape locally adapted genomes in spite of homogenising gene flow. In this chapter I have taken advantage of recently developed genomic resources in *D. montana* (Parker et al., *in prep*) and the extensive resources in other *Drosophila* species to conduct a population genomic analysis of populations at different latitudes from North America and Finland.

In general, patterns of genetic diversity are similar across the populations with Finnish populations showing slightly lower nucleotide diversity (π) overall than North American populations. This pattern is consistent with a recent population bottleneck and subsequent expansion in Finland (Mirol et al., 2007). In general the patterns of nucleotide diversity across the populations are consistent with recent estimates from mitochondrial DNA. Finnish and Canadian populations have lower diversity than populations from further south in N.A. (Mirol et al., 2007). Mitochondrial data also indicated a smaller N_e in among Finnish populations, consistent with lower nucleotide diversity (Mirol et al., 2007).

The one outlier is the high altitude population at Crested Butte which shows very

low π and strongly negative values of Tajima's *D* overall. Negative values of Tajima's *D* are often interpreted as evidence for selective sweeps or background selection (Huber & Lohmueller 2016). However, neither of these processes should reduce these statistics throughout the genome. An alternative explanation is a recent bottleneck and population contraction followed by a recovery period. It is unclear whether populations in Crested Butte survive the entire winter where they were collected or whether they overwinter further down the slope and recolonise/expand into new habitat every spring. Such oscillations might produce local populations which are depleted in genetic diversity. Thus inference about selective sweeps in this population should be made with caution and further population genomic sampling will be required at different altitudes through successive years to determine whether this is the case.

Nevertheless, many SNPs show an association between genetic differentiation among populations and environmental differentiation across the environmental variables in PC1 and PC2 from a PCA of climatic data. This study shows that chromosome four contains a disproportionate number of thes SNPs. Interestingly, other studies show that chromosome four harbours QTL for several ecologically relevant phenotypes in *D. montana* (Tyukmaeva et al., 2015). One explanation for this result is strong selection and high rates of hitchhiking resulting in many SNPs with similar allele frequency patterns across populations. Another alternative is that co-localised variants in this region are contributing to local adpatation. *D. montana* as a species is polymorphic for many inversions, including on chromosome four (Stone 1960; Morales-Hojas et al., 2007). These inversions show patterns of fixation and polymorphism across populations that suggest they are not driven by neutral drift (Morales-Hojas et al., 2007). However, these inversions and their clinal distributions remain understudied.

Inversions have often been found to vary clinally and contribute to differentiation throughout the cline (e.g. Kolaczkowski et al., 2011; Cheng et al., 2012; Kapun et al., 2016). Such findings are consistent with a more important role for inversions in adaptation than previously thought (Hoffman & Riesberg 2008). Other traits that delineate different populations of this species and may, to some extent, contribute to reproductive isolation between them (Jennings et al., 2014), also localise to regions with known polymorphic inversions (Schäfer et al., 2010; Schäfer et al., 2011; Lagisz et al., 2012). Taken together, our lack of knowledge about the distribution

inversion polymorphisms in this species is a crucial gap in our understanding of the patterns of genomic variation, and represents an area in which results would yield fruitful further insights in the geographic distribution of alleles and phenotypes as well as the forces that drive them.

The recent efforts to sequence the genome of *D. montana* also surveyed the rates of molecular evolution in cold tolerant and non-cold tolerant species of *Drosophila*. The genes found to be evolving at faster rates in cold-tolerant species were found to be enriched for many of the same functional categories as in this study (e.g. Leucine-rich repeat, Glyco- and Lipo-proteins, membrane proteins, ion transporters; Parker et al., *in prep.*). This suggests that the same types of pathways involved in conferring greater cold-tolerance on different species are under selection in current populations in response to climatic stressors. For example, ion transport and homeostasis is an important part of cold tolerance adaptations across *Drosophila* (MacMillan et al., 2015a; 2015b). In particular, the hemolymph concentrations of Na^+ and K^+ are correlated with cold tolerance (MacMillan et al., 2015a). Similarly, membrane proteins and lipids are an important determinant of membrane and cuticular permeability at different temperatures, which in turn has an effect on the resistance to desiccation stress in insects (Gibbs 2002; Stanziano et al., 2015). There is also evidence for a close link between the desiccation stress response and cold tolerance across species and in *Drosophila* in particular, suggesting an overlap in some of the pathways involved (Sinclair et al., 2007). These results indicate that some of the same biochemical processes that are being targeted by selection on larger scales (across species; Parker et al., *in prep.*) are also involved in local adaptation for different populations within a species. Further confirmation of this trend would provide a nice link between "micro-" and "macro-" evolutionary processes**.**

Adaptations involved in cold adaptation and diapausing behaviours in *D. montana* seem to be intricately linked (Vesala et al., 2011). Diapause depends on the photoperiod calendar which senses the change in photoperiod throughout the seasons and induces developmental changes (Koštál, 2011). A related phenomenon is the circadian clock which has been a topic of much study in *Drosophila* (Schlichting et al., 2016; Hellfrich-Förster 2017). The extent to which these systems are linked by a common molecular mechanism remains unknown but the available evidence points to a

role for canonical "clock" genes (like *tim*) in sensing the potoperiod and changing seasons (Koštál, 2011). The biochemical function of the circadian clock involves both negative and positive feedback loops of transcription factors which in conjunction with other tissue specific transcription factors contribute to rhythmicity in other physiological processes (Hellfrich-Förster 2017). Meanwhile the entrainment, the regulation of the rhythms with respect to the outside environment, is strongly dependent on the photoperiod (light/dark cycles) and reception/transduction of these signals by the compound eye (Shlichting et al., 2016; Hellfrich-Förster 2017) but also of daily temperature cycling (Hellfrich-Förster 2017). In this context, I highlight a few genes which lie near SNPs that show an association between genetic and environmental differentiation among populations.

Photoreception and Entrainment of the Circadian Clock

Photoreception in insects occurs via the compound eye but also via light sensitive organs called "eyelets" (Hellfrich-Förster et al., 2002; Hellfrich-Förster 2017). In *glass/sine oculis* double mutants, which do not have functional "eyelets", or compound eyes, the entrainement of the circadian clock is defective (Hellfrich-Förster et al., 2002; Hellfrich-Förster 2017). The genes *glass* and *sine oculis* are important transcription factors which are necessary for the correct patterning of compound eyes and eyelets as well as the expression of phototransduction proteins (Hellfrich-Förster et al., 2002). As such they play a role in building the organs which transmit information about the external environment to the internal circadian clock. Meanwhile the gene *inaC* (or eye-PKC) is an important mediator of the visual signalling pathway (Wang et al., 2008). Thus, several genes important in the formation and function of organs involved in the entrainment and photoperiod dependent cycling of circadian rhythms are linked to SNPs that show an association with environmental variables.

Circadian Rhythms

The genes *timeless* (*tim*), *Clock* (*clk*), *supernumerary limbs* (*slmb*), *vrille* (*vri*), *timeout, aurora borealis* (*bora*), and *discs overgrown* (*dco*) all occur near SNPs which are significantly associated with environmental variation among the populations in this study. These are all related to the function of the circadian clock in various ways. *Clk* is

central to the circadian rhythm system and regulates several physiological processes (Hellfrich-Förster 2017). *tim* functions as a mediator between external light environment and inhibition of different parts of *per/clk* cycling (Hellfrich-Förster 2017). Variants of *tim* are known to have a latitudinal clinal distribution in European *D. melanogaster* with higher frequencies in the south (Tauber et al., 2008; Pegoraro et al., 2017), the opposite pattern is seen in populations in the eastern United States (Pegoraro et al., 2017). One variant of this gene (*ls-tim*) shows faster rates of entry to diapause in response to the photoperiod (Tauber et al., 2008; Pegoraro et al., 2017). However, European clines in the diapausing phenotype are not related in a simple manner to the frequencies of the *ls-tim* allele highlighting the polygenic nature of a complex trait like diapausing behaviour as well as the effect that novel mutations have on adaptive phenotypic clines (Pegoraro et al., 2017). Nevertheless, it is a prime candidate gene at the interface of the outside light environment and internal circadian clocks.

A recent study of diapausing *D. melanogaster* found transcriptional responses in several genes. In particular, both *tim* and *bora* (*aurora borealis*) were differentially expressed in flies diapausing for 3 weeks compared to 1 week old flies (Kuçerova et al., 2016). A study of activity rhythms and *per*/*tim* expression cycling in *D. montana* revealed that expression cycling is broadly similar between diapausing and non-diapausing flies within the same photoperiod regime (Kauranen et al., 2016). Overall, *tim* and *per* showed very different patterns of cycling compared to *D. melanogaster* in similar photoperiod treatments (Kauranen et al., 2016). These results further highlights the need to investigate the expression patterns of these genes before and during diapause entry in flies from different populations (Kauranen et al., 2016).

Parker et al., (2015a) observed differential expression of *vri* in response to cold acclimation, as well as several other genes involved in the regulation of the circadian clock. *vri* is a repressor of *clk* transcription and thereby contributes to *clk* cycling (Hellfrich-Förster 2017). In another study, *dco* and *slmb* were shown to have lower expression in diapausing than reproductively active females (Kankare et al., 2010). Although *slmb* was also downregulated in older non-diapausing flies compared to young non-diapausing flies suggesting that age and not diapause might be driving changes in transcription. *Slmb* is involved in the binding of phosphorylated *per* and movement to the proteasome for degradation (Hellfrich-Förster 2017). Meanwhile, *dco*

(also called DBT) is also involved in the binding and degradation of *per* (Hellfrich-Förster 2017) and is listed with several mutant phenotypes in FlyBase including locomotor and eclosion rhythms.

Cold Acclimation and Cold Tolerance

In a recent study, diapausing and non-diapausing flies were also investigated for transcriptional differences in *D. montana* with several genes showing transcriptional differences between flies kept in different light cycles, diapausing or non-diapausing flies, as well as their interaction (Parker et al., 2015). In *D. montana* an RNA-seq study was carried out to identify transcriptional responses to cold-shock and transcriptional changes during cold-acclimation. Several previously identified genes were observed to change expression levels (Parker et al., 2015). One novel gene was *Inos,* which produces the protein *myo*-inositol-1-phosphate synthase, a part of the inositol synthesis pathway (Parker et al., 2015). Knockdown RNAi experiments confirmed that reducing the expression of *Inos* leads to higher rates of mortality when exposed to cold (Vigoder et al., 2016). Another example is *DnaJ-1* (also called Hsp40) which is a heat shock chaperone protein constitutively expressed in *Drosophila* (Neal et al., 2006; Colinet et al., 2010). *DnaJ-1* was observed to be upregulated in response to heat stress in *D. melanogaster* (Neal et al., 2006). It also shows lower levels of expression during cold stress in *D. melanogaster,* followed by an increase in expression during recovery (Colinet et al., 2010). Similarly, in response to cold shock *DnaJ-1* is upregulated in the Collembolan *Folsomia candida* (Waagner et al., 2013). This gene is downregulated in response to cold acclimation in *D. montana* (Vesala et al., 2012c) Other heat shock proteins (Hsps) have also been found to show changes in expression during cold stress indicating a wider role for them in temperature stress responses that vary across taxa (Vesala et al., 2012c).

In sum, many genes are involved in the form and function of the general components of photoperiod dependent circadian clock. Undoubtedly many complicated interactions between genes play a role and there are still gaps in our understanding. Nevertheless, several of the genes that in this study are linked to SNPs showing an assocation with environmental variation are directly involved in these processes and therefore present prime candidates for further study. Follow up work should strive to

use technologies such as RNA interference (RNAi) or CRISPR/C*as9* (e.g. Vigoder et al., 2016) to validate these and other genes for the contribution of variants to variation in cold tolerance and diapause response.

Studying species that show clinal distributions of traits can help our understanding the forces that drives population divergence and local adaptation. Perhaps the best studied species with clinal variation in traits is *Drosophila melanogaster* (Adrion et al., 2015). Clines in cold-tolerance, body and wing size among other phenotypes are known from North America and Australia (e.g. Kolaczkowski et al., 2011; Machado et al., 2015; Adrion et al., 2015). Study of these clines has revealed some evidence of adaptive clines in individual gene alleles (e.g. Schmidt et al., 2008) and inversions (Kapun et al., 2016). However, there are also cautionary notes that clinal genetic variation can also be produced by largely demographic processes. For example, Machado et al., (2016) found that genomic clines in *D. simulans* are not as stable across seasons as those in *D. melanogaster* and that isolation by distance patterns were not as strong in *D. simulans.* The authors suggested that clines in *D. simulans* could be the result of strong bottlenecks in winter at higher latitudes (resulting in low diversity at high latitudes) and subsequent summer gene flow from further south (Machado et al., 2016). Similar results are seen in this study for the high altitude population at Crested Butte which shows extremely low diversity throughout the genome. Clearly further work is needed to assess the contributions of demographic processes to the maintenance of the geographic patterns in genetic diversity also in *D. montana*.

Other studies have cast additional complexity even on the *D. melanogaster* clines in North America and Australia. These populations are very recent colonisations and there is now evidence to suggest there has been recent admixture from ancestrally European and African populations in both contintents that contribute to some of the clinal patterns in allele frequencies (Bergland et al., 2016). These types of demographic effects on pattern is less likely in *D. montana* because current distributions of this species are likely much older (Mirol et al., 2007).

Finally, this study has provided some methodological insights as well. The observation that quasibinomial GLMs do not seem to behave well in this scenario is a useful result. This is clear from the distribution of p-values which do not follow the expected distributions (either a uniform distribution or one skewed toward lower p-

values). This is probably due to a combination of factors. First correlated p-values at closely linked SNPs may produce an abundance of p-values of a particular class. Second non-linear relationships between the allele frequencies and latitude or environmental differentiation will produce spurious p-values if a linear regression is performed. Regardless, it is clear that p-value distributions are not as expected and thus correction for multiple testing by standard methods (here q-values) produce dubious results. The methods employed in BayeScEnv perform much better to uncover associations. Finally, another potential approach which may perform well is to apply the beta-binomial Gaussian process (BBGP) method of Topa et al., (2015). While this was developed to study time-series experimental evolution data in theory it should apply equally well to continuous variables like latitude.

5.5 Concluding Remarks

Studying clines of populations that vary in ecologically important phenotypes will give insights into population divergence and speciation. With the advent of next generation sequencing it has become possible to investigate the loci that underlie variation in interesting traits and to study how genetic differentiation progresses. In this chapter I took advantage of clines of the frigophilic fruitfly *D. montana*. This species shows variation among populations in ecologically important traits like cold tolerance and the diapause response. Meanwhile, the ever decreasing costs of next generation sequencing has allowed the development of genomic resources in a draft genome and accompanying annotation as well as several transcriptome studies to identify candidate genes involved in these traits. In this chapter I took a population genomic approach to sequence pools of individuals from populations sampled from extremes of the *D. montana* range. I used PC analysis to determine the main axes of environmental variation among these populations. I then used recently developed Bayesian methods to test for an association between genetic differentiation among populations and the environmental differentiation along the environmental PC axes. Many SNPs show a significant association and these cluster throughout the genome. The non-random distribution of significant SNPs throughout the genome also hint at the strong co-localisation of candidate loci possibly within inversions which are known to segregate

in natural populations of *D. montana*. Finally, many genes in close linkage to these significant SNPs have been identified as candidate genes in previous gene expression studies within a few populations. Patterns of genetic diversity around many of these genes were suggestive of recent selective sweeps or background selection in some or all of the populations.

Chapter 6 Comparative genomics of crows and signals of positive selection in the genome of the New Caledonian crow (*Corvus moneduloides*).

Abstract

Comparative genomics is a powerful approach to understanding the forces that drive evolution at the level of genes. Studying the genomes of ecologically similar species can give insights into the evolutionary processes affecting different loci. Identifying signatures of selection within genomes of different species can give insights into the genetic loci that are important in producing adaptive differences between species. Particular genes or gene families might be under selection during the colonisation of and adaptation to new environments. Alternatively non-coding regions might be under selection which can be detected by uncovering regions of low diversity which are indicative of selective sweeps. In this chapter I take advantage of a multi-species sequencing effort of the genus *Corvus* to identify signatures of selection within the NC crow.

New Caledonian (NC) crows are of interest due to their tool-using behaviour. The recent discovery of tool use in another species of crow (the Hawai'ian crow) allows an opportunity to understand the conditions that favour the evolution of such behaviours. Many species of crows, including the NC and Hawai'ian crows, are also island endemics. This means they have likely experienced very particular demographic histories, including population contractions and adaptations toward island habitats. The availability of multiple island species allows a control separating the effects of island colonisation from specific selective forces shaping the NC crow. In this project I was primarily interested in identifying signatures of selection throughout the genome of the New Caledonian (NC) and Hawai'ian crows. Taking advantage of a large number of newly generated sequences for several crow species I investigate coding sequence evolution, population genetic differentiation, and signatures of selective sweeps. The aim was to identify genomic regions and loci that show differences between the NC

crow and other species due to selection.

I find that few coding sequences show robust evidence of positive selection within the NC crow lineage. Those that do are associated with promising functions and pathologies (e.g. maintenance of attention, schizophrenia and general intelligence) in humans and mice. Additionally, a number of regions show reductions in diversity (Tajima's *D,* Fay and Wu's *H*) that lie outside the range expected from evolution by neutral drift. Some of the genes within these regions are known to be involved in the development of beak morphology.

Author Contributions

This chapter is part of a collaborative project involving multiple research groups and as such many other people have contributed to data presented in this chapter in the following ways. Sampling of blood and feather samples for DNA extraction and the DNA extraction itself was carried out by the members of Christian Rutz research group (University of St Andrews, Scotland), the Jochen Wolf research group (Ludwig-Maximilians-Universität Műnchen, Germany), and the Robert C. Fleischer research group (University of California, Santa Barbara, U.S.A.). Advice on the construction of the phylogenetic tree was given by Darren J. Parker (University of Lausanne, Switzerland). Additionally, much of the methods, including pipelines, were developed through discussion between Verena Kutschera (Uppsala Universitet, Sweden), Nicolas Dussex (University of Otago, New Zealand) and myself. With the above acknowledgements, all analyses in this chapter are my own work.

6.1 Introduction

6.1.1. Comparative Genomics

Evolutionary biology is necessarily comparative. Understanding the forces that produce the diversity of organisms requires a careful comparison of their characteristics and how they vary with ecology. Since the advent of large scale, low cost next-generation sequencing comparative studies have been extended to comparative genomics (Ellegren 2008; 2014; Pardo-Diaz et al., 2015). This involves the study of genomic variation across populations and species that differ in important characteristics

in order to understand the forces that give rise to these differences. It is now possible to sequence entire genomes of multiple individuals even in non-model organisms. Patterns of variation within the genome allow inferences about the demographic history of a species or population, locations of recent or ongoing selective sweeps (e.g. coding sequences evolution, or gene family expansions and contractions), and structural rearrangements (inversions, transpositions, etc.). Inference is made on the basis of population genetic theory which makes predictions about the patterns that are expected under different evolutionary scenarios (Charlesworth & Charlesworth 2008; Huber & Lohmueller 2016). Comparative genomics, by identifying genomic regions showing signatures of selection or different demographic patterns, can compare these patterns of variation across species or populations which have different evolutionary and ecological histories to better understand the genetic differences between populations. (Ellegren 2008; Ellegren et al., 2014; Pardo-Diaz et al., 2015). Thus, rather than identifying and accounting for every variant that affects a phenotype, comparative genomics, in effect, focuses on those loci and regions that contribute to differences between species or populations.

Many studies are proving successful in identifying loci producing adaptive difference between populations and species. In sticklebacks (*Gasterosteus aculeatus*), freshwater colonisation is characterised by the repeated loss of specific traits (including bony, armor plates; Hohenlohe et al., 2010; Jones et al., 2012). Multi-population sampling (using reduced representation sequencing methods, RAD-seq) and the development of a draft genome studies showed that differentiation between pairs of marine and freshwater populations was localised to the region containing the *Eda* locus (Hohenlohe et al., 2010). Further sampling using individual whole-genome sequencing show that parallel colonisations of freshwater systems are characterised by selection at the same variants present in the global marine population (Jones et al., 2012). The locations of these regions showing evidence of recurrent adaptive sweeps corresponded very closely to early work identifying the causal locus. QTL methods and positional cloning had previously identified the *Eda* locus which was associated with the loss of armor plates (Colosimo et al., 2004; Colosimo et al., 2005). These loci do not constitute a complete accounting of the loci that contribute to variation in skeletal morphology within sticklebacks, but they do identify some of the loci which are important in

producing ecologically adaptive differences between populations. It is also possible to show that variants of these loci are under selection in the new environment (Barrett et al., 2008) and as such they inform us about the process of evolution acting on genetic variants.

A recent example involving colour patterns in eurasian crows also highlights the power of this approach. The eurasian *C. corone* species complex is characterised by several subspecies with different colour patterns that come into contact at multiple hybrid zones (Poelstra et al., 2014; Vijay et al., 2016). With large scale population sampling of genomes researchers identified regions that were characterised by strong differentiation (F_{ST}) across these hybrid zones and reduced amounts of diversity with populations indicative of selective sweeps (Poelstra et al., 2014; Vijay et al., 2016). These regions contained candidate genes that have roles in the production and deposition of melanin such as *CACNG4* and *CACNG1* (Poelstra et al., 2014). Some candidates even co-localise closely with genes involved in transduction of signals from the eyes to the brain as well as opioid and dopamine signalling (Poelstra et al., 2014). These results raise the possibility of a mechanism of cosegregation of preference and trait genes in this system (Poelstra et al., 2014).

Clearly these genes do not explain all of the variation in plumage colour or mate choice behaviour within these species but they are strong candidates for the genes and variants that play an important role in the differentiation of these subspecies or populations in the face of gene flow. Thus they represent evolutionarily important loci. Many other examples of the identification of important loci that contribute to differences between populations are available (Stinchcombe & Hoekstra 2008; Ellegren 2014). In summary, comparative genomics is a very productive approach towards identifying the loci or genomic regions that differ between populations and species. Analytical methods can also identify regions that are likely to have been under selection in the divergence of populations. Thus, comparative genomics can help researchers identify loci that contribute to adaptive differences between populations and species.

6.1.2 Avian Comparative Genomics

A wealth of genomic resources is becoming available for many avian lineages (Balakrishnan et al., 2010; Ellegren et al., 2012; Ellegren 2013; Wolf et al., 2014;

Zhang et al., 2014; Lamichhaney et al., 2015). Birds show an enormous diversity in lifestyles, ecological history, mating systems and other characters. There are patterns emerging from the study of avian genomes (Ellegren 2007; 2013). For example, in birds females are the heterogametic sex (ZW) in contrast to mammals where males are the heterogametic sex (XY), thus theories of sex chromosome evolution can be compared. Just as the X chromosome in other systems has a lower N_e than autosomes, resulting in lower diversity, N_e for the Z chromosome in birds is lower. This also has the effect of reducing diversity on Z relative to autosomes in general (Ellegren 2013). Differences in mating system, which alters the N_e of the Z chromsomes, also affects Z:autosome diversity ratio (Corl & Ellegren 2012; Ellegren 2013). In addition, chromosomal organisation in birds is highly conserved in comparison to other taxa. Chromosome within a genome vary in sizes, overall recombination rates, and in the density of coding sequence (Ellegren 2007; 2013). In diverse lineages there is a relationship between nucleotide diversity, chromosome length, and recombination rates which highlight the importance of genome organisation in determining patterns of variation (Ellegren 2007; 2013). This accumulation of data presents an excellent opportunity to conduct comparative studies of adaptations to different environments to identify the action of selection within diverging genomes. Expectations of pattern can be derived from results (e.g. higher differentiation on the Z chromosome, and a negative correlation between diversity and chromosome length) and be controlled for in the study of new species. In this chapter I consider the corvid radiation and, in particular, the New Caledonian (NC) crow (*C. moneduloides*), a tropical island species, which has attracted much attention over the years for its tool use in foraging.

6.1.2 The New Caledonian Crow

Crows in general, and the NC crow in particular, are increasingly recognised for their cognitive abilities (Emery & Clayton, 2004). The manufacture and use of tools in the wild by NC crows contributes to this view (Emery & Clayton, 2004; Jønsson et al., 2012). In the wild, NC crows make different types of tools that differ in their sophistication (e.g. the number of steps required to produce them, and the specificity of the materials used) and the amount of manufacturing required to produce them (Emery & Clayton, 2004; Rutz & St Clair, 2012). The characteristics of these tools vary both

temporally and geographically, suggesting a "cultural" aspect to tool design, although some environmental factors also show an effect on the distribution of different tool types (Hunt & Gray 2003). Indeed, studies on the ontogenetic development of tools in the crows show an effect of social learning (Kenward et al., 2006). Although relatively little is known about the foraging functions/benefits of these tool types evidence suggests that they are employed in order to gain access to nutritionally superior food sources (Rutz & St Clair, 2012). These studies have prompted questions about the general ecological conditions that favour tool use behaviour in the NC crow.

A diversity of relatively closely related species exist within the corvid radiation (e.g. Jønsson et al., 2012; Häring et al., 2012) with many differences in ecology and life style. They also have some convergent colonisations of similar habitats which, presumably, have exerted similar evolutionary pressues. The NC crow, a tropical island species, has a closely related sister species, *C. woodfordi,* which is native to the Solomon islands, also tropical islands. Thus a kind of ecological "control" or contrast species exists for comparative genomics.

Some evidence strongly indicates a genetic predisposition toward aspects of tool use (Kenward et al., 2005; 2006). Naïve juveniles raised in captivity have been observed to fashion rudimentary stick tools in order to solve foraging tasks (Kenward et al., 2005, 2006). Tool use is a complex trait (in genetic terms) and variation in this trait stems from variation in many other underlying traits. Other morphological and behavioural traits, also likely heritable, are thought to be adaptations that facilitate the use tool use and its development in NC crows. For example, changes to bill and skull morphology giving straighter bills and more binocular vision are indicative of adaptations to the unique load distribution and precision required for tool use (Troscianko et al., 2012; Matsui et al., 2016). Some evidence also points to NC crows having larger brains, relative to body size, than other crows (Cnotka et al., 2008; Jønsson et al., 2012). NC crows also share derived brain structures called perineural glial clusters with other passerine birds, analogous to structures that are found in humans and mice (Medina et al., 2013). Behavioural traits in the NC crow also include persistent object exploration (Holzhaider et al., 2010; Kenward et al., 2011). Many of these and other traits are not categorically unique to NC crows but may be common adaptations associated with tropical island colonisation. There is now strong evidence to suggest that tool use

developed independently also in the Hawai'ian crow (Rutz et al., 2016). The Hawai'ian crow also has uncharacteristically straight bills and a similar ecology and habitat to the NC crow (Rutz et al., 2012; Rutz et al., 2016). Although the species is extinct in the wild a captive breeding population exists and a study shows that naïve juveniles develop tool use on their own (Rutz et al., 2016).

A wealth of data exists on the genetic basis of morphological traits in birds from the zebra finch (*Taeniopygia guttata*; Warren et al., 2010), chicken (*Gallus gallus*; Abzhanov & Tabin 2004), and not least from the many species that comprise Darwin's finches (Abzhanov et al., 2004, 2006 ; Lamichhaney et al., 2015). Also, the pathways involved in craniofacial development are highly conserved across birds and other vertebrates (Brugmann et al., 2010; Bhullar et al., 2015). Behavioural traits are less well studied. These data give us a reasonable foundation on which to build hypotheses about the types of loci that may be under selection in NC and Hawai'ian crows as a cause or consequence of tool-using behaviour.

In this study, I use comparative genomics to compare coding sequence evolution across several species of corvids. We might expect elevated rates of evolution (nucleotide substitutions) in candidate genes involved in the determination of beak/skull morphology and shape. We might further expect elevations in the rate of evolution in these genes within both the NC crow and Hawai'ian crow lineage if there has been convergent evolution at important loci. However, the data available for the Hawai'ian crow are limited (with only one individual sampled), making strong inferences about signatures of selection within this species difficult. Other changes in response to the colonisation of novel habitats might be changes to gene repertoire sizes as some classes of genes become more important or redundant (McBride 2007; Gardiner et al., 2008; Cortesi et al., 2015). I also conduct a genome-wide survey of diversity in the genome to assess the signals of selective sweeps. While the focus is on the NC crow lineage, the results also bear on the genomics of island species, and the effects of non-selective demographic forces on genomic patterns of diversity. Population contractions and expansions are expected to change patterns of diversity throughout the genome while selection should alter patterns locally around favoured loci. Using another species with a similar evolutionary history but differing in the traits of interest should function as an ecological control. Thus, regions underlying adaptively important changes unique to the

NC crow should be characterised by signatures of selection (e.g. reduced diversity, and elevated rates of amino acid substitutions) in the NC crow but not in the closely related *C. woodfordi* which is not known to use tools.

6.2 Materials and Methods

6.2.1 Sampling and Sequencing

Sampling and DNA extraction was carried out by various methods and people (see *Author Contributions*). Sequencing was performed on multiple platforms at various sequencing centers. Some of the genomes use in this study have been previously published (Poelstra et al., 2014; Vijay et al., 2016). Original reads from these genomes were obtained and processed in the same way as outlined for novel sequencing runs below. The species and the number of individuals sequenced are given in table 6.1.

As part of recent work to survey the amount of repetitive sequence in *Corvus corone cornix* (hereafter *C. cornix*) scaffolds of the draft genome were ordered onto hypothetical chromosomes through a synteny based approach (Weissensteiner et al., 2017). Individual chromosomes from several bird species were independently aligned to the *C. cornix* reference genome. Scaffolds were ordered by a principle of parsimony such that if two outgroups shared an ordering it was considered ancestral and used. In the case of unresolved ordering, the order according to alignment to the chicken was used (Weissensteiner et al., 2016). Here I assume the same scaffold ordering information for the purpose of genome-wide window scans.

6.2.2 Mapping and Consensus Genome Building

In order to build consensus genomes for each species in this study, sequenced reads were mapped to the *C. cornix* reference genome following a standard procedure. First reads were trimmed to remove any potential TruSeq2 or TruSeq3 paired end adapters using Trimmomatic (0.32) (Bolger et al., 2014). Next, each sample was mapped separately with bwa mem (0.7.13) (Li 2013). Alignments with a mapping quality < 10 were removed with samtools (v.1.3) (Li et al., 2009). Duplicate reads/alignments were removed from the files with samtools. To build a consensus genome for each species variants are first called using samtools mpileup and bcftools (v.

1.3) (Li et al., 2009) with all available .bam files for each species to create a .vcf file for each species. Indels are then removed with GATK (v 3.4.0) (McKenna et al., 2010) to conserve reference annotation coordinates. This constricts consensus genome length to be the same in all species but allows the use of the *C. cornix* annotation coordinates to extract coding sequences. Variant sites within species were treated in one of two ways; either they were completely masked (replaced by "N" and ignored in downstream analyses) from the consensus genomes to keep only fixed sites. Alternatively, variant sites were filtered using vcftools (v. 0.1.14) (Danecek et al., 2011) to keep only sites where the non-reference allele had a frequency greater than 0.5. The rationale is that variants which are under selection and on their way to fixation should be at higher frequencies than the reference allele at a given site. Thus, if any site in the reference genome had a non-reference (alternative) allele with frequency greater than 0.5 in the other crow species the alternative allele replaced the reference allele in the consensus genome. This procedure is less conservative than masking all variant sites but probably more conservative than considering the alternative allele regardless of its allele frequency. Consensus genomes were created using the bcftools "consensus" routine with default parameters. Finally, sites with a coverage < 5x were masked from consensus genome. To accomplish this, first all .bam files across individuals for a species are merged with bamtools (v. 2.3.0) (Barnett et al., 2011). Total coverage achieved at each site was then computed with bedtools (v. 2.25.0) (Quinland & Hall 2010) to produce a .bed coordinate file of all sites with a coverage < 5x. This coordinate file was then used to mask the consensus genomes with bedtools.

Because the species *C. corone orientalis, C. corone corone,* and *C. cornix* share a substantial amount of variation (Vijay et al., 2016) a three species *C. cornix* species group was created. Five individuals from each of *C. orientalis, C. corone*, and *C. cornix* were chosen to represent the Eurasian spread of this species. The .bam files with the most mapped reads to the *C. cornix* reference genome from each individual were chosen and a consensus genome for the three species *C. corone* group was built as described above. For additional population genetic parameters, .bam files from the same individual were merged for the five *C. moneduloides* individuals and the five *C. woodfordi* individuals.

6.2.3. Core Genes and Molecular Phylogenetics

The genome completeness assessment tool BUSCO (v 1.22; Simão et al., 2015) was run for each consensus genome using the set of 3,023 vertebrate core genes (Simão et al., 2015). The coding regions of the set of common completely recovered genes across the 12 *Corvus* species and *T. guttata* were extracted with gffread (cufflinks v 2.2.1; Trapnell et al., 2010; Roberts et al., 2011) from each consensus genome. Any sequences with more than 20% Ns were removed from the set. The final set contained 860 coding sequences. These sequences were individually codon aligned using PRANK (v. 100802) (Löytynoja & Goldman 2005). All alignments were then concatenated into a single alignment. A likelihood tree was computed using RaxML (Stamatakis 2014) using a GTR + gamma model of sequence evolution with 4 rate categories for each codon position. A tree was produced using maximum likelihood and 10,000 bootstrap iterations with *T. guttata* specified as the outgroup.

6.2.4 Ortholog Discovery

Analysis of coding sequence evolution depends on robust identification of orthologous sequences in the species compared. To this end, protein sequences for two independently annotated species (*C. cornix*, and *T. guttata*) were compared. Where more than one CDS occurred for a single gene (i.e. multiple annotated isoforms) the longest CDS was kept, except where otherwise noted. These anotations consist of the newest annotation for *C. cornix* (RefSeq: GCF_000738735.1, release 100) available from NCBI (Poelstra et al., 2014; NCBI 2016) and the latest releast of the *T. guttata* annotation from ENSEMBL (v.3.2.4; release 87.1; Cunningham et al., 2015). In total, the annotations included 19,456 sequences from *T. guttata* (from 16,377 unique CDSs), and 24,318 sequences from *C. cornix* (from 14,149 unique CDSs).

To determine the orthologous relationships between the CDSs in the *C. cornix* and *T. guttata* annotations the two annotations were compared using a reciprocal BLAST (Camacho *et al.*, 2009) and OrthAgogue (v 1.0.2) (Ekseth *et al.*, 2014) to find the best reciprocal hits for each CDS. Only the longest CDSs (16,377 CDSs in *T. guttata* and 14,149 from *C. cornix*) were used. Any sequences that did not have a single best hit, e.g. a *T. guttata* sequence that had equally good hits to multiple *C. cornix*

sequences or *vice versa*, were treated as unreliable because a good ortholog could not be established, these were discarded.

6.2.5 Assessing Positive Selection Along the NC Crow Lineage

Nucleotide Substitution Rates – PAML Analysis

To assess the evidence for positive selection among coding sequences in the NC crow, different models of nucleotide substitution rates were compared using the branch-models in codeml from the PAML (v.4.6) package (Yang 2007). Sequences where the proportion of "Ns" in the sequence was > 0.2 were discarded. Additionally, each coding sequence from across the crows must have an unambiguously identified ortholog in zebra finch (see *Ortholog Discovery* above). To determine how robust results were to the choice of species in the analysis, several comparisons were made which included different numbers of species of crow (5 species, 7 species, and 8 species). Across the 5 crow species set (*C. moneduloides*, *C. dauuricus*, *C. frugilegus*, *C. splendens* and the 3 species "*C. corone*" group, and *T. guttata* as an outgroup) 10,102 coding sequences filled these criteria and were kept for analysis. The 7 species analysis used the same species as in the 5 species set with the addition of *C. tasmanicus* and *C. corax*, resulting in 9,944 coding sequences. Finally, the 8 species set included all the species in the 7 species group with the addition of *C. hawaiiensis,* giving 9,934 coding sequences. Translated sequences were re-aligned in PRANK (v. 150803) (Löytynoja & Goldman 2005).

Branch-models assume a single *dN/dS* ratio (ω) for the entire sequence which can vary across branches of the phylogeny. For the 5 and 7 species sets two models of sequence evolution were considered. The "null" model assumed a common ω across the entire phylogeny (model = 0, NSsites = 0, fix_omega = 0). The "alternative" model assumed one ω for the branch leading to the NC crow and a different one for all the other branches (model = 2, NSsites = 0, fix_omega = 0). A comparison of these models tests whether there is a different rate of sequence evolution on the branch leading to the NC crow compared to the rest of the phylogeny. For the 8 species set including *C. hawaiiensis* a number of alternative models were run. One model considered a strict convergent evolution scenario wherein the "null" model was as described above but in the "alternative" model the branches leading to the NC crow and to the Hawaiian crow

shared the same ω (all other parameters were as above). A separate model is also considered which allows different ω values for the NC crow and Hawaiian crow lineages.

The null and alternative models were compared by a likelihood ratio test (LRT) in R (v 3.3.1; R Development Core Team, 2016) with 1 d.f. All p-values are converted to q-values (Storey 2002; Storey & Tibshirani 2003) using the R package *qvalue* (v. 2.4.2) (Storey et al., 2015) to control the false discovery rate (FDR) at a threshhold of 0.05**.**

6.2.6 Gene Family Expansions and Contractions

Gene gain and loss among the three independently annotated species (*C. moneduloides*, *C. cornix*, *T. guttata*) was assessed using the software CAFE (v. 3.1) (de Bie et al., 2006; Hahn et al., 2005). CAFE estimates gene family contractions and expansions by first estimating the probability of gene gain/loss (lambda) using the data from all gene families (Hahn et al., 2005). This global estimate of lambda is then used to parameterise a birth/death model of gene family evolution which can be used to obtain a distribution of likelihoods for each gene family over a range of family sizes at the root.

Gene family sizes were obtained by building ortholog groups using OrthoFinder (v. 0.6.0) (Emms et al., 2015). Following recommendations, only the longest CDSs for each gene from *C. cornix* and *T. guttata* were used (de Bie et al., 2006; Hahn et al., 2005). Since data on the number of alternative transcripts (CDSs) per gene are unavailable for *C. moneduloides*, all CDSs were used. The CDSs were also filtered to exclude very short, dubious sequences (< 10 amino acids). In OrthoFinder (v 0.6.0) the default options were used to cluster genes into ortholog groups. The output of OrthoFinder was formatted for input to CAFE in R by counting the number of genes in each ortholog group for each species. The gene family data were also filtered to include only families with at least one gene in every species.

The phylogenetic tree (see above) was pruned using the R package *ape* (v. 3.5) (Paradis et al., 2004) to contain only the three species (*C. cornix, C. moneduloides* and *T. guttata*) under analysis. CAFE requires an ultrametric tree with branch lengths corresponding to units of time since the last common ancestor. Substitution rates were

converted to units of time using the chronos() function in the R package *ape* the smoothing parameter "lambda" was set to 0, which allows substitution rates to vary completely across the branches (Paradis et al., 2004), an estimated date of divergence between *C. cornix* and *C. moneduloides* was taken as 10-11 million years from Jønsson *et al.,* (2012) and the confidence interval of (36-50) of divergence time between *C. corone* and *T. guttata* from TimeTree (Hedges et al., 2006) to calibrate the tree. Finally, CAFE was run with default parameters and lambda was estimated from the data.

6.2.7 Population Genetic Parameters and Divergence

To identify genomic regions that differed strongly between the NC crow and *C. woodfordi* I compared the "landscapes" of diversity and divergence in these species. Several population genomic statistics of diversity and differentiation were calculated. Genetic diversity is quantified by π, Tajima's *D*, and Fay and Wu's *H* as calculated from the site frequency spectrum (SFS) in ANGSD (Korneliussen et al., 2014). These statistics (Tajima's *D* and Fay and Wu's *H*) contrast different estimates of nucleotide diversity with that of Watterson's θ. The different estimates are more sensitive to the contributions of low frequency alleles (Tajima's *D*) or high frequency derived alleles (Fay and Wu's *H*). Large discrepancies between them and Watterson's θ are taken as evidence of deviations from neutrality (e.g. selective sweeps, or population expansions/contractions; Huber & Lohmueller 2016).

Estimation of Fay and Wu's *H* requires a full SFS and an ancestral genome to polarise variants as ancestral or derived (Korneliussen et al., 2014). To this end, an "ancestral" genome was built using three outgroups to the NC crow (*C. cornix*, *C. dauuricus* and *C. frugilegus*). The read alignments mapped from all of these species were combined using bamtools merge. A consensus genome was then called using all the information from these species. This ensures that only sites which are fixed in all three species are kept in a final consensus genome as the ancestral state, variable sites are treated as unreliable and masked. This procedure is similar to what other studies have done in ancestral genome reconstruction (Poelstra et al., 2014; Vijay et al., 2016). Statistics were calculated for 50kb sliding windows (with a step size of 10kb) on the aligned scaffolds. These windows and scaffolds were then placed on the *in silico* chromosomes of Weissensteiner *et al.,* (2016) to assess genome-wide patterns.

Since these population genetic statistics have no intrinsic distribution that is independent from parameters (such as N_e, and the mutation rate) hypothetical neutral population, significance threshholds can only be derived by simulation (Huber & Lohmueller 2016). To this end, I performed simple population genetic simulations, as in previous chapters. I use analytical solutions to population genetic equations that describe allele frequency changes in populations under the neutral model (Charlesworth & Charlesworth 2008). These solutions describe the distribution of allele frequencies in a population at mutation-drift equilibrium (Charlesworth & Charlesworth 2008). This distribution is a beta distribution B(α, β), where;

$$\alpha = 4N_e v$$

and;

$$\beta = 4N_e u$$

where u and v are the forward and reverse mutation rates (per site per year). SNPs are simulated by first sampling the allele frequency (p_A) from this distribution. N diploid genotypes are then produced by sampling from a multinomial distribution where the probabilities are given by assuming Hardy-Weinberg equilibrium ($p^2 + 2pq + q^2$). The allele count A is then $p^2 + pq$. The SFS is thus built up assuming N sampled individuals from a population. Finally, population genetic statistics π and Tajima's D are estimated from this SFS according to standard equations (Tajima 1989; Korneliussen 2013).

Because the effective population size of NC crows and *C. woodfordi* is unkown, I use a range of values for N_e (1,000, 10,000, 20,000, 100,000, 200,000, 1,000,000, and 2,00,000). Though previous estimates of population sizes are between 100,000 and 1,000,000 (Ellegren 2007), both *C. woodfordi* and NC crows are relatively restricted island species which means the true population size is likely lower. Similarly, a range of mutation rates have also been reported for different bird lineages (Nam et al., 2010; Ellegren 2013; Smeds et al., 2016), ranging from 1.91×10^{-9} in chicken (Nam et al., 2010) to 2.3×10^{-9} in Ficedula flycatchers (Smeds et al., 2016), and phylogenetic studies indicate variation across lineages (e.g. Lanfear et al., 2010). Thus I run simulations over the range 1.23, 1.91, 2.21 and 2.3×10^{-9}. Because sample sizes in this study are quite low (five individuals for the NC crow and *C. woodfordi*) I also use a range of values for N (5, 10, 50, and 100) to examine how the sample size affects estimation of these population genetic statistics. For each combination of parameters I simulate 100 site

frequency spectra of 50,000 SNPs and obtain distributions for Tajima's *D*. These distributions are used to set a threshhold for identifying windows with low Tajima's *D* in the genome (see above). Many of steps in the above methods rely on in-house scripts. An archive of pipeline descriptions and R scripts used in the analysis and plotting of data is available at (http://github.org/RAWWiberg/ThCh6).

6.3 Results

6.3.1 Mapping and Consensus Genome Building

General patterns of base composition were very similar across all species (table 6.1), although genome size was constrained to be the same (indels are ignored) to conserve the coordinates of annotated coding sequences from *C. cornix*. GC content is highly conserved across all species and only marginally higher in the *C. cornix* reference genome. BUSCO gene set analysis suggests similar overall levels of completeness across all the crows probably owing to the relatively good quality draft reference genome of *C. cornix* used in this study. The levels of completeness are overall similar to those seen in *T. guttata* (figure 6.1).

6.3.3 Molecular Phylogenetic Tree

The phylogenetic tree produced in this study largely corroborates previous hypotheses (Häring et al., 2012; Jønsson et al., 2012; Rutz et al., 2016). The genetic distance separating the NC crow lineage from the base of the Corvids is only 0.004 substitutions per site (figure 6.2). This indicates very low levels of genetic divergence among the crows. The branch lengths are much lower than other studies, probably because this chapter does not use mitochondrial genes. The topology also corroborates conclusions in recent studies that tool use likely arose independently in *C. moneduloides* and *C. hawaiiensis*. There is one obvious difference, the placement of *C. frugilegus* with *C. kubaryi* and *C. splendens* rather than with *C. hawaiiensis* as in Häring *et al.,* (2012) and Jønsson *et al.,* (2012).

Table 6.1. Summary statistics on the consensus genomes and the reference genome (R). Shown is the total length of the genomes in Mb, the counts (in millions) of A, T, G, and C. N (missing sites or coverage < 5x) , and the GC content. Also given is the number of individuals included in the study in square brackets.

Species	***Length***	***N (%)***	***A***	***T***	***G***	***C***	***GC (%)***
Corvus corone cornix (R)	1049.97	0.03	299.60	298.20	212.80	212.40	40.50
C. corone	1049.97	0.04	296.30	296.33	209.96	210.32	40.03
species group [15]							
C. brachyrhynchos [6]	1049.97	0.04	294.76	294.61	208.23	208.48	39.69
C. corax [1]	1049.97	0.03	297.20	297.02	211.05	211.27	40.22
C. dauuricus [4]	1049.97	0.04	294.74	294.54	208.07	208.30	39.66
C. frugilegus [4]	1049.97	0.04	295.06	294.89	208.91	209.14	39.81
C. hawaiiensis [1]	1049.97	0.03	297.27	297.09	210.95	211.17	40.20
C. kubaryi [1]	1049.97	0.03	297.62	297.44	211.65	211.87	40.34
C. monedula [4]	1049.97	0.04	295.62	295.43	209.14	209.36	39.86
C. moneduloides [5]	1049.97	0.03	297.30	297.12	211.23	211.44	40.26
C. splendens [5]	1049.97	0.04	296.16	295.99	210.16	210.38	40.05
C. tasmanicus [1]	1049.97	0.03	297.34	297.16	211.25	211.47	40.26
C. woodfordi [5]	1049.97	0.03	297.50	297.32	211.61	211.82	40.33

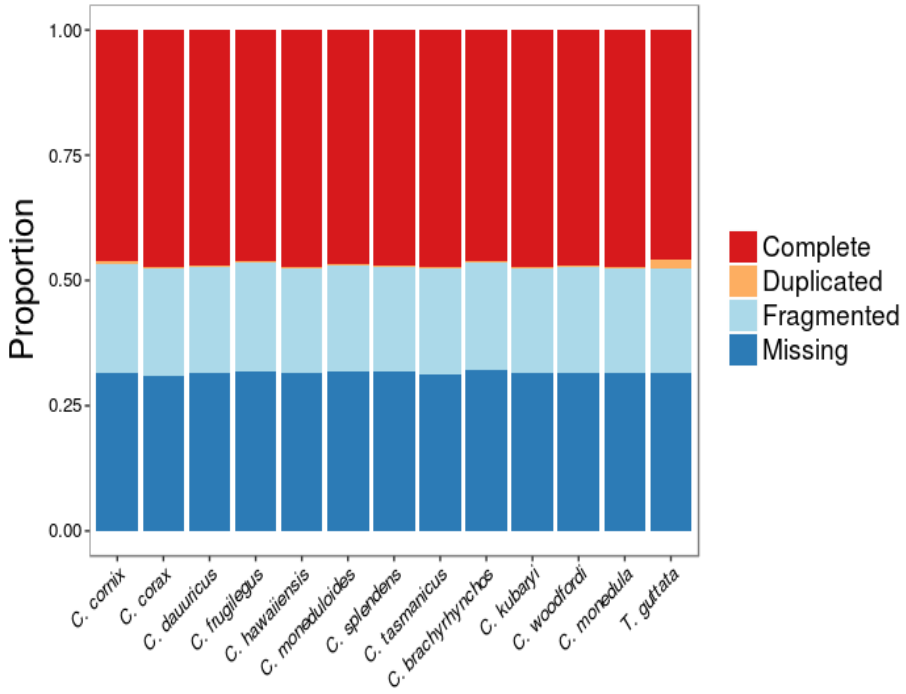


Figure 6.1. The proportions of complete, complete (but duplicated), fragmented, and missing genes from the BUSCO gene completeness analysis.

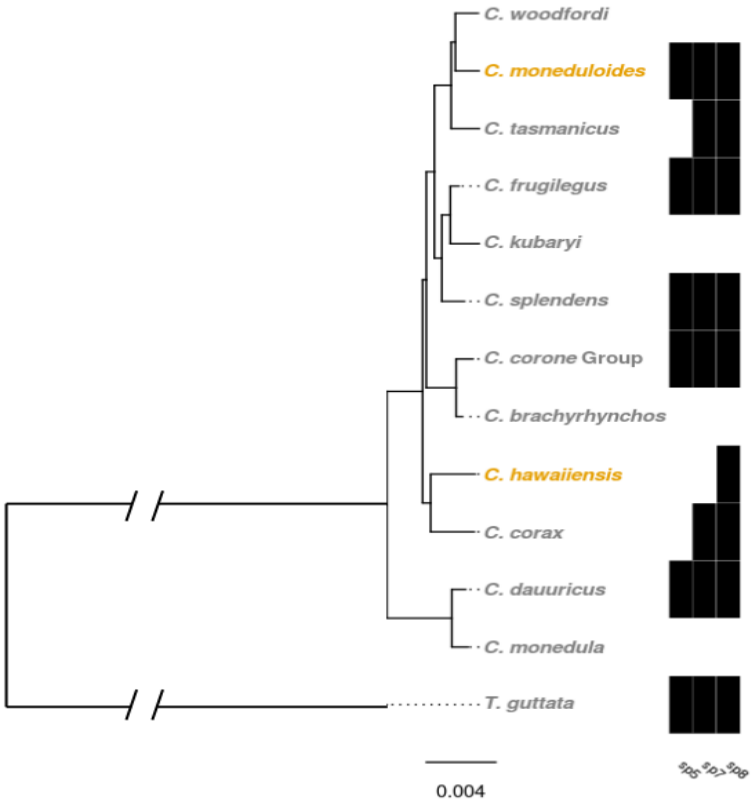


Figure 6.2. Maximum likelihood phylogenetic tree of 12 *Corvus* species based on concatenated alignments of 860 core vertebrate genes. Species shown in orange are known for tool use. Black boxes to the right of the tree indicate which species were used in phylogenetic sequence evolution analyses. Branch lengths are proportional to the number of substitutions (see scale bar). Species names have been aligned to the right and linked to the tip of their branches with dashed lines.

6.3.5 Rates of Molecular and Gene Family Evolution

The number of genes for which there is evidence of a different rate of molecular evolution in the NC crow lineage when compared to the other crow lineages depends on which set of species is chosen (table 6.2). Between ~19 and 54 genes in all species sets show evidence for a different rate of molecular evolution in the NC crow lineage. Across all species sets, 171 unique genes show different rates of evolution in NC crow.

Most of the genes which show a different rate of evolution in the NC crow lineage have a $\omega < 1$ (table 6.2).

Most of these genes show support for the alternative model in only one or two species sets. Only 10 genes are supported by the alternative model in more than 3 species sets and only two of these have $\omega > 1$ indicating positive selection within the crow lineage. The gene *Dtnbp1* shows differential rates of molecular evolution in the NC crow lineage compared to the other crow lineages in five sets. Furthermore, *dN/dS* ratios (ω) are > 1 (ranging between 1.5 – 1.8) in each case except one. The only case where ω is not > 1 is in the analysis which assumes the same rate for the NC and Hawai'ian crow lineages (strict convergence). By contrast *Gse1* and *Srsf1* seem to be consistently conserved in the NC crow lineage in comparison to the other crow lineages with ω close to 0. *Htt* shows elevated rates evolution in three sets with ω of between 100 and 999 which is driven by almost no synonymous changes in the sequence and a few non-synonymous changes. Removing poorly aligned regions with Gblocks (v. 0.91b; Castresana 2000) and re-running the analysis for *Htt* does not change the result ($\omega > 999$; $p < 0.001$). However, in the case of *Dtnbp1* the sequence is substantially truncated by Gblocks and the result is lost because the region containing a single non-synonymous substitution in the NC crow is removed. Other genes show $\omega > 1$ but only in a few of the species sets. However, many of these results have very high ω values which may be the result of alignment or annotation assembly error.

The Hawai'ian crow lineage was only tested for different rates of evolution in the 8 species sets. Two analyses tested a different ω in the Hawai'ian crow lineage compared to the NC crow lineage. These analyses differed only in whether variant sites were included in the consensus genome or not (see *6.2 Methods*). Across these two datasets there were no genes with a $\omega > 1$ in both sets.

Finally, Gene family contraction and expansion analysis shows evidence of 104 and 140 expansions and contractions respectively for the *C. cornix* lineage. In contrast, there are 297 and 194 expansions and contractions respectively for the *C. moneduloides* lineage. However, there are only 14 and 517 expansions and contractions, respectively, at the base of the *Corvus* clade. Many of these expansions and contractions are in extremely large gene families of uncharacterised genes in all three species. However, a few more modest effects are seen as well.

Table 6.2 The results from PAML analyses for different species sets. Shown are the number of genes for which null and alternative models were tested, the number of genes which were significant for the alternative model, and the number of genes with $\omega >$ or $<$ 1. Sequences either contained variant sites or variant sites were masked (see *6.2 Methods*). For the 8 species sets, the number of genes with w > or < 1 are given for NC crow and Hawai'ian crows separated by a "/".

Species set	**N Genes (N significant)**	**N (ω >1)**	**N (ω < 1)**
5 species set[a]	10,102 (23)	4	19
5 species set[a] (masked)	10,102 (18)	4	14
7 species set[b]	9,944 (19)	7	12
7 species set[b] (masked)	9,944 (22)	6	16
8 species set[c]	9,934 (21)	5/4	16/17
8 species set[c] (masked)	9,930 (52)	12/8	47/51
8 species set[c2]	9,934 (17)	4/4	13/13
8 species set[c2] (masked)	9,930 (38)	7/7	31/31

a. *C. moneduloides*, *C. dauuricus*, *C. frugilegus*, *C. splendens,* the *C. corone* species group, and *T. guttata*

b. *C. moneduloides*, *C. dauuricus*, *C. frugilegus*, *C. splendens,* the *C. corone* species group, *C. tasmanicus, C. corax* and *T. guttata*

c. *C. moneduloides*, *C. dauuricus*, *C. frugilegus*, *C. splendens,* the *C. corone* species group, *C. tasmanicus, C. corax, C. hawaiiensis*, and *T. guttata*

2. The alternative model assumes strict convergence between the NC crow and the Hawai'ian crow (see *6.2 Methods*).

6.3.6 Population Genetic Parameters and Divergence

In both the NC crow and *C. woodfordi* chromosome-wide mean π is correlated with chromosome length. Shorter chromosomes have higher diversity than longer chromosomes (figure 6.3**)**. If chromosome-wide levels of π are treated as independendent estimates, the correlation between π and chromosome length is highly significant in both species (Spearman rank correlations, NC crow: rho = -0.71, S= 7676, $p < 0.001$ *C. woodfordi*: rho = -0.89, S = 8494, $p < 0.001$). Estimates of π, Tajima's *D*,

Fay and Wu's *H* as well as F_{ST} are given in figure 6.4 and figure 6.5. Overall levels of diversity are substantially lower in *C. woodfordi* than in the NC crow (table 6.3; figure 6.4). Mean values of Tajima's *D* are also slightly above 0 in both species but more so in *C. woodfordi*. Estimates of π are in general much lower in *C. woodfordi* than in the NC crow (table 6.2; figure 6.4).

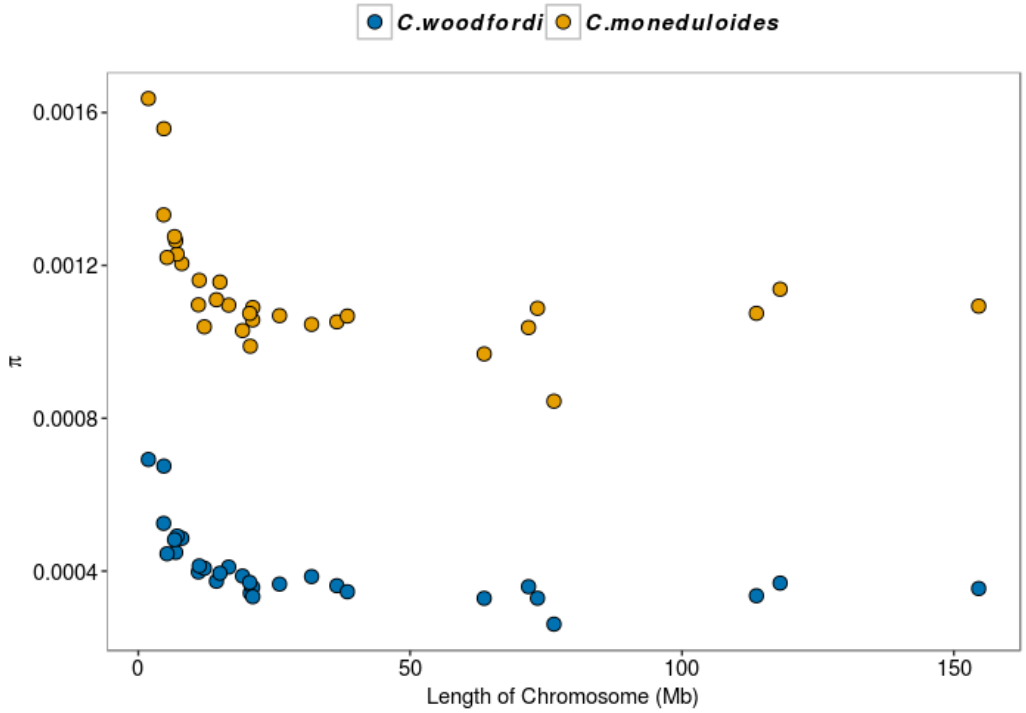


Figure 6.3. The relationship between chromosome length and mean estimates of nucleotide diversity (π) in the NC crow (*C. moneduloides*) and *C. woodfordi*.

F_{ST} between the two species is quite high overall, and is highest on the Z chromosome, as expected (figure 6.5). F_{ST} also shows a strong relationship with π in both species (figure 6.6). Low diversity regions have much higher F_{ST} and a few peaks of diversity are associated with troughs of F_{ST}.

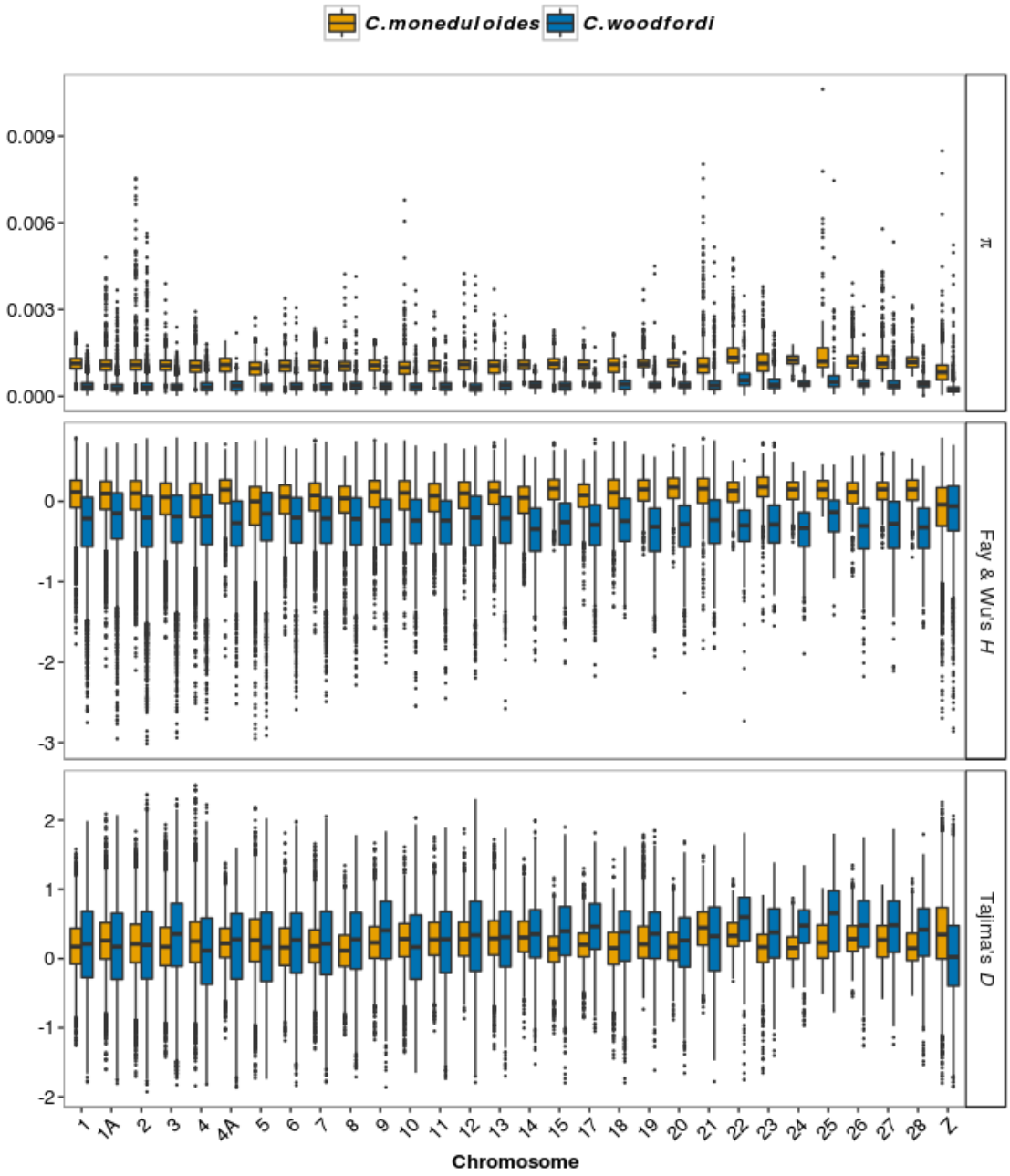


Figure 6.4. Estimates of π, Fay and Wu's *H*, and Tajima's *D*. The data shown are from ~100,000 windows spread across all chromosomes.

Table 6.3. Summaries of population genetic statistics in the NC crow and *C. woodfordi.* Shown are the mean, median (in brackets), and range (in square brackets) of each statistic across the genome. Results are split by Autosomes (A) and the Z chromosome (Z). Also given is the A:Z ratio of diversity (π).

	C. moneduloides			**C. woodfordi**		
	A	**Z**	**Z:A**	**A**	**Z**	**Z:A**
π ($x10^{-3}$)	1.1 (1.1)	0.85 (0.83)	0.77	0.36 (0.32)	0.26 (0.22)	0.72
	[0.13, 10.6]	[0.042, 8.5]		[0.017, 7.5]	[0.030,5.2]	
Tajima's D	0.22 (0.21)	0.36 (0.35)	-	0.23 (0.26)	0.042 (0.024)	-
	[-1.84, 2.50]	[-1.80, 2.26]		[-1.93, 2.37]	[-1.85, 2.06]	
Fay and Wu's H	0.031 (0.088)	-0.11 (-0.042)	-	-0.28 (-0.22)	-0.13 (-0.063)	-
	[-2.95, 0.79]	[-2.70, 0.79]		[-3.02, 0.79]	[-2.86, 0.70]	

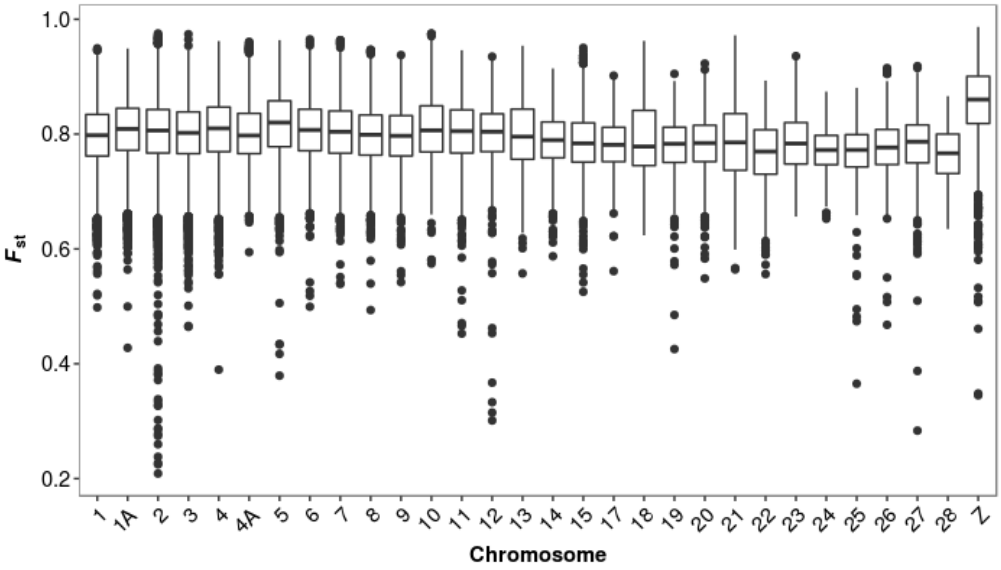


Figure 6.5. F_{ST} calculated between *C. woodfordi* and the NC crow. The data shown are from ~100,000 windows spread across all chromosomes.

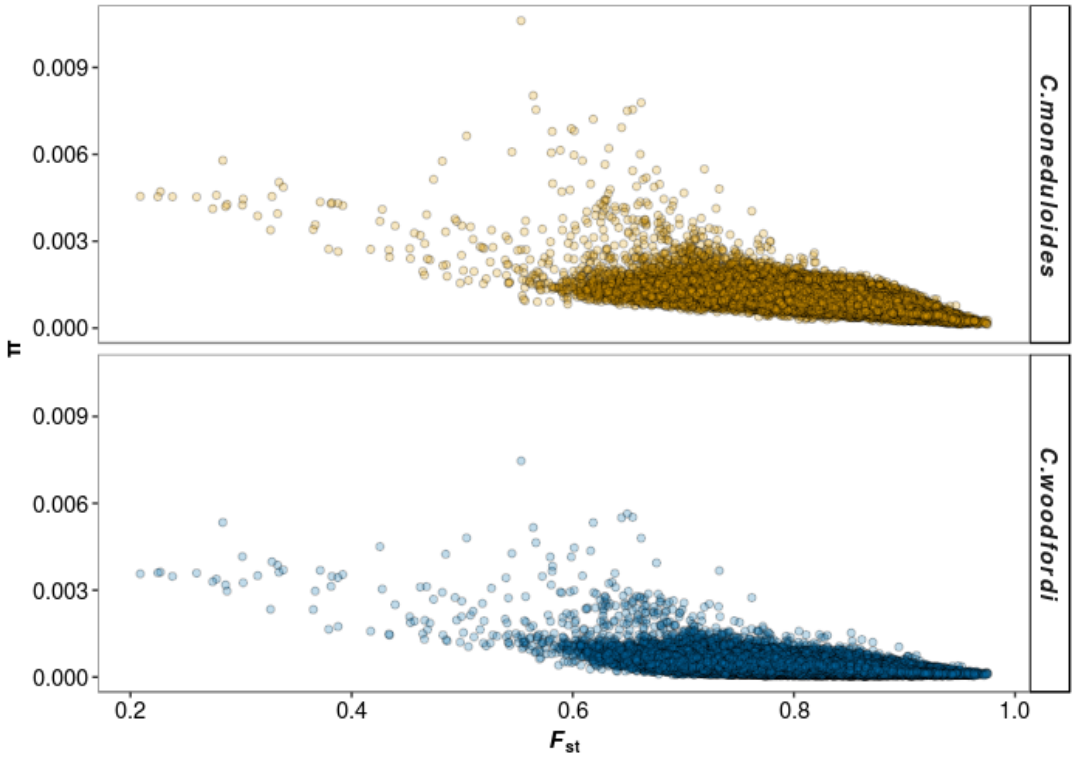


Figure 6.6. The relationship between nucleotide diversity (π) and genetic differentiation (F_{ST}) across windows in the NC crow (*C. moneduloides*) and *C. woodfordi*.

Population genomic simulations show some variation in the expected levels of Tajima's *D* under different assumptions of the effective population size, mutation rates and the number of sampled individuals (figure 6.7). The biggest source of variation is N_e and so inferences about regions showing evidence of selective sweeps will depend mostly on what is a realistic estimate of N_e in *C. woodfordi* and the NC crow. The 1st and 5th percentiles of the distributions of Tajima's *D* range from -1.6 to -0.05 (table 6.4) and -1.3 to -0.07 (not shown) across the parameter combinations respectively. Sample sizes primarily affect the variance in the distributions at lower N_e (figure 6.7). A conservative threshhold for identifying "significant" windows is therefore set at -1.6. Windows with values of Tajima's *D* below this are considered candidate regions undergoing a selective sweep. In total, 19 windows show values of Tajmia's *D* lower than -1.6 in the NC crow. Some of these windows are adjacent to one another. These can be combined into 11 1Mb regions across different chromosomes for further investigation (figure 6.8 and 6.9).

Table 6.4. The range, across different values of the mutation rate, of the 1st percentile of the distribution of Tajima's *D*. Values are shown for different sample sizes (*N*) from populations with different N_e.

	***N* = 5**	***N* = 10**	***N* = 50**	***N* = 100**
***N_e* = 1,000**	-1.48, -1.40	-1.60, -1.44	-1.62, -1.06	-1.44, -1.18
***N_e* = 10,000**	-1.40, -0.90	-1.19, -0.89	-1.55, -1.08	-1.23, -0.91
***N_e* = 20,000**	-1.01, -0.79	-0.91, -0.84	-1.26, -0.75	-1.11, -0.79
***N_e* = 100,000**	-0.46, -0.31	-0.44, -0.34	-0.55, -0.38	-0.56, -0.35
***N_e* = 200,000**	-0.41, -0.28	-0.49, -0.25	-0.37, -0.22	-0.37, -0.25
***N_e* = 1,000,000**	-0.15, -0.09	-0.15, -0.09	-0.22, -0.08	-0.18, -0.08
***N_e* = 2,000,000**	-0.13, -0.07	-0.09, -0.05	-0.08, -0.03	-0.09, 0.01

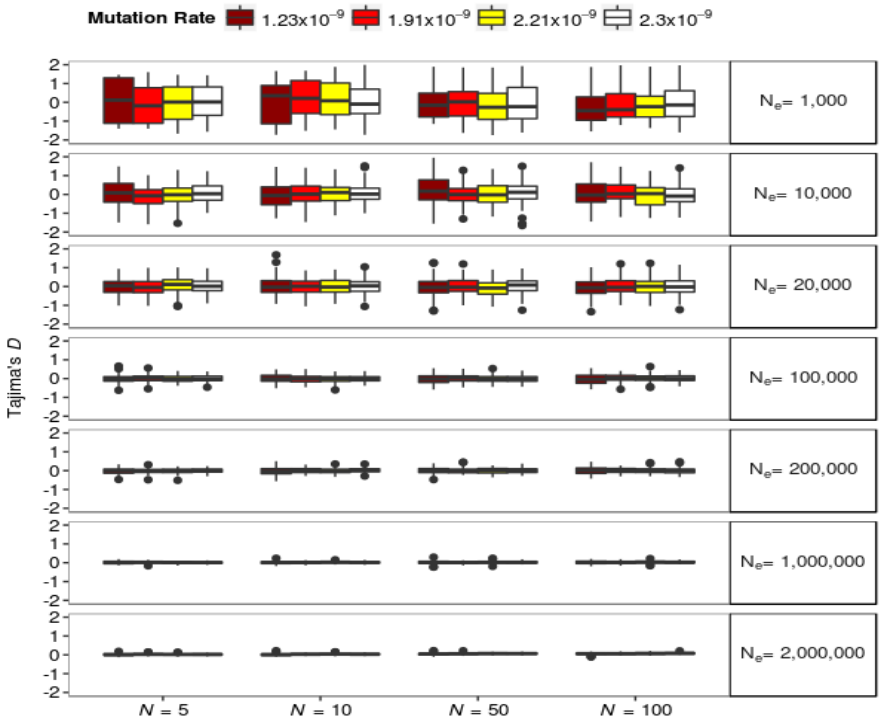


Figure 6.7. Results from population genetic simulations. Shown are the distributions of Tajima's *D* obtained under different assumptions of N_e, mutation rates, and the number of sampled individuals (*N*). Points at the ends of whiskers are outliers.

In some of the regions in figures 6.8 and 6.9, Tajima's *D* is reduced in both *C. woodfordi* and the NC crow while in others the reductions are more obvious in the NC crow (figure 6.8). For example, the region between 26.8 and 29 Mb on chromosome five shows an extended region of reduced Tajima's *D* in the NC crow while in *C. woodfordi* this region is closer to neutral expectations (figure 6.8). These trends are more apparent for Fay and Wu's *H* (figure 6.9). In contrast, none of these regions show any obvious peaks or troughs in F_{ST} (not shown).

These regions contain a total of 350 genes, though many are uncharacterised genes that have no known ortholog in the zebra finch. Of the 261 that do have such an ortholog, 11 are known to be expressed in the zebra finch brain according to ZeBRA (ZeBRA: A Zebra Finch Expression Atlas, http://www.zebrafinchatlas.org). If the mouse orthologs of these genes are submitted to the MamPhEA tool (Weng & Liao 2010), which uses information from the mammalian phenotype ontology database (Smith et al., 2004), several phenotypic categories are marginally significantly enriched (though not after Bonferroni correction) including "small maxilla", and "abnormal cerebellar lobule formation". Still other genes in these regions are known to be involved in the development of beak and skull morphologies from other birds (*Dkk2*, and *Calm1*).

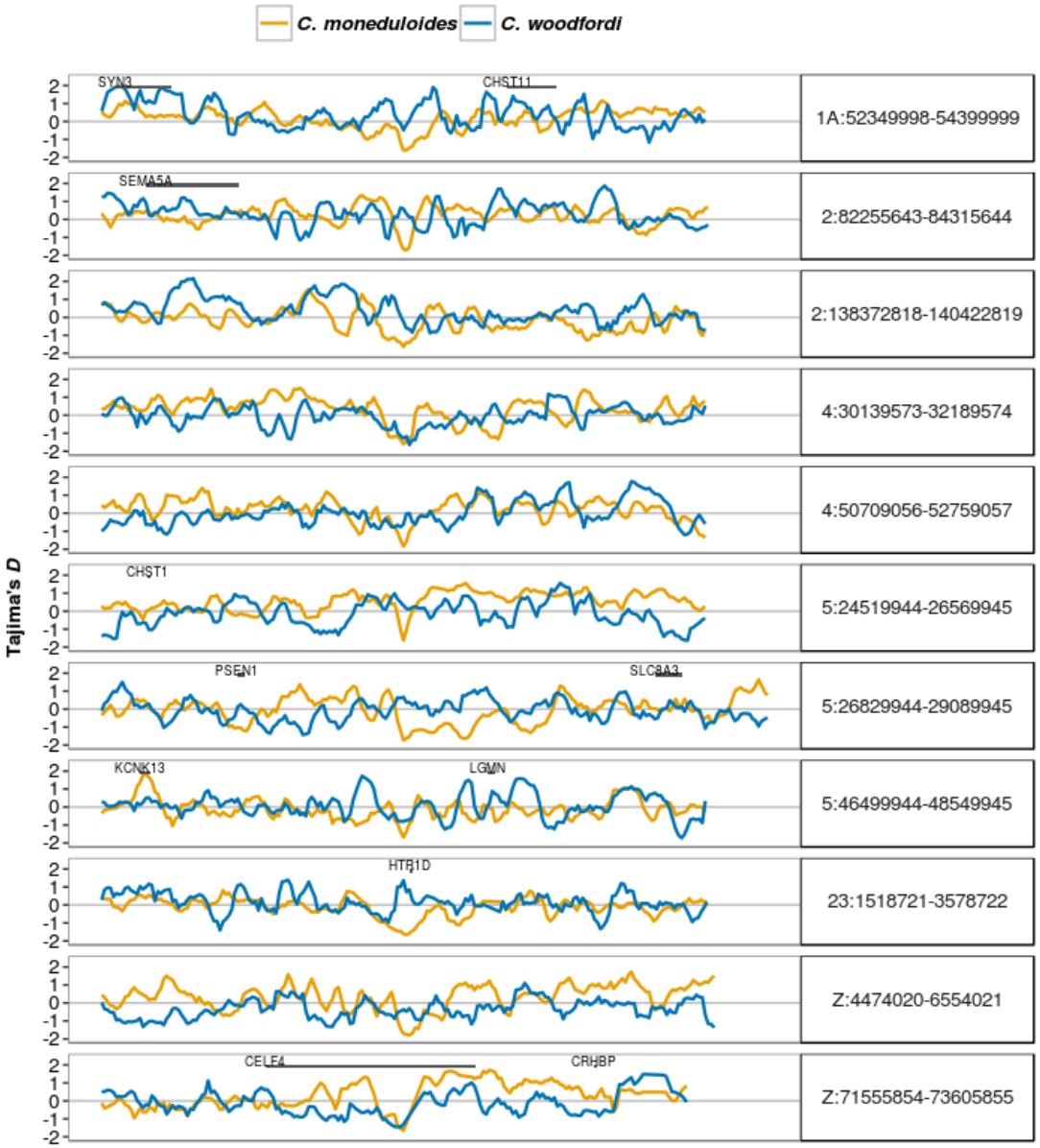


Figure 6.8. Tajima's *D* within the regions up to 1Mb around windows with Tajima's *D* reduced. Data are shown for sliding windows of 50kb with a step size of 10kb. Also shown are the locations of the 11 genes within these regions which are expressed in the zebra finch brain.

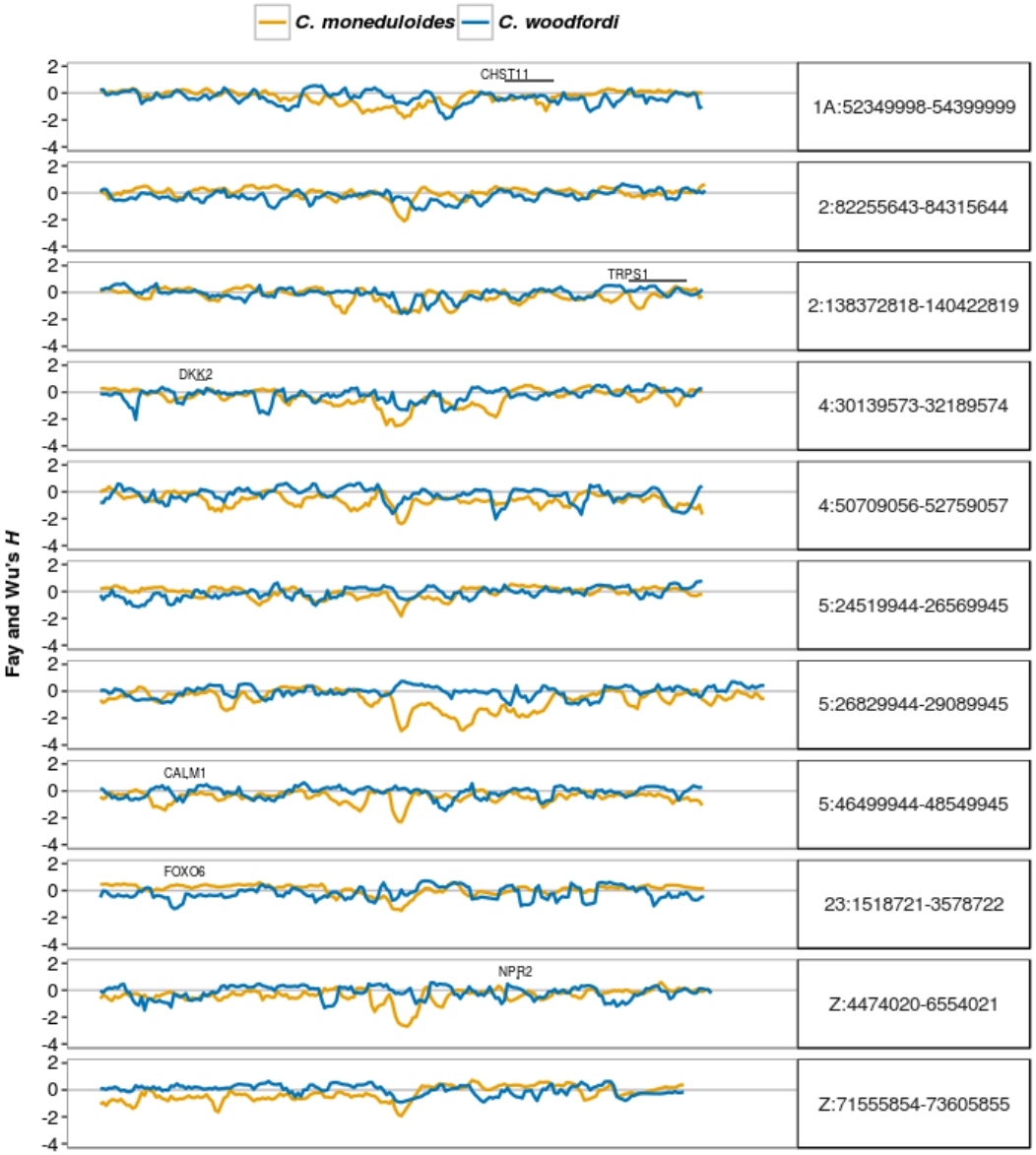


Figure 6.9. Fay and Wu's *H* the regions around windows with reduced Tajima's *D*. Data are shown for sliding windows of 50kb with a step size of 10kb. Also shown are the six genes within these regions which have been associated with beak and skull development or are classified in the "small maxilla" mouse phenotype category.

6.4 Discussion

6.4.1 Coding Sequence Evolution

A comparative genomic approach can help to understand the processes that give rise to differences between populations and species. It can be useful in identifying the loci involved in producing important differences between species. This chapter takes a comparative genomic approach to look for signatures of selection throughout the genome of the NC crow. Although the Hawai'ian crow is included in some analyses, the relative paucity of data for this species and the closest relatives means that strong inferences about species differences are difficult to make, therefore I focus on the NC crow. I take two main approaches; first I ask if there is any evidence of positive selection within coding sequences along the NC crow lineage. I also take advantage of genome sequences for multiple samples from two closel y related tropical island species to assess patterns of sequence diversity around the genome and uncover signatures consistent with recent selection within the NC crow.

The results from an analysis of rates of molecular evolution within coding regions suggests that there is little evidence for diversifying selection on coding sequences within the NC crow lineage. Only two genes are relatively consistently identified as being under positive selection. These two genes are good candidates for follow up. *Dtnbp1* (*dystrobrevin binding protein 1*) is a gene in which variants have been associated with schizophrenia in human patients in multiple studies (Sabb et al., 2009; Cheah et al., 2015; Bakanidze et al., 2016) and within schizophrenic patients variants affect other schizophrenia associated phenotypes (e.g. "sustained attention", "set-shifting", and "hallucinations"; Cheah et al., 2015; Bakinadze et al., 2016). There is also some evidence for a role of variants of *Dtnbp1* in "intelligence" among the general, healthy, human population (Sabb et al., 2009). Meanwhile, mutations of *Htt* (Huntingtin) cause Huntington's disease symptoms which include reduced attention and memory (Saudou & Humber 2016). In addition, heterozygous mouse mutants of the same gene show motor and cognitive defecits (homozygotes do not survive embryonic development; Nasir et al., 1995). These phenotypes are interesting because some of the cognitive traits thought to be particular to NC crows include extraordinary persistence and attention in foraging tasks as well as the potential for a socially learned aspect to

particular tool designs (Holzhaider et al., 2010; Kenward et al., 2011).

The lack of strong evidence for convergent evolution in coding sequence between the NC and Hawai'ian crows is perhaps not surprising. Similar phenotypic results could very well be obtained through changes at non-coding loci. For example, in other birds it seems that changes in transcription of key genes during development determine beak shape and morphology (Abzhanov et al., 2004a; Wu et al., 2006; Brugmann et al., 2010; Mallarino et al., 2011). Given this, changes in different loci could alter transcription in similar ways to produce the same phenotypes in different species. This is also the conclusion of a recent study of convergent molecular evolution of marine mammal adaptations (Foote et al., 2015).

That very few coding sequences seem to be under strong positive selection within the NC crow lineage in comparison to the other crow lineages is perhaps also unsurprising. Many studies have indicated that transcriptional changes are the source of variation across species in many adaptations, rather than coding sequence changes (but see Rands et al., 2013). For example differences in beak shapes across many species of birds seem to be driven by a few candidate genes but it is primarily differences in the timing of their expression that matter and not coding sequence (Abzhanov et al., 2004; 2006; Abzhanov & Tabin 2004). Cartilage outgrowth during beak development in chicken is driven by differential expression of *Fgf8* and *Shh* during embryonic development (Abzhanov & Tabin 2004b). Meanwhile, shape is controlled by a few conserved regulatory pathways involving *Bmp4, Dkk3*, *Igf2r* and others (Wu et al., 2006; Brugmann et al., 2010; Mallarino et al., 2011).

Other studies suggest it is also the timing and locations of expression rather than absolute levels of expression that matter for beak shape (Wu et al., 2006). Thus, causal variants may lie outside the coding region and will not be discovered by analyses which consider only coding sequence molecular evolution. These results bear on the debate about the relative importance of *cis*-regulatory regions *versus* structural (protein) coding changes in adaptive evolution (King & Wilson 1975; Carroll 2005). This debate has its roots in a discussion of the discoveries in the 1960s and -70s that coding sequences of chimpanzees and humans were virtually identical. Whence, then, the obvious differences between the species (King & Wilson 1975)? It has since been argued that regulatory regions are the key. These regions are more likely spots for adaptive changes

to occur because downstream pleiotropic effects are minimised (Carroll 2005; Stern & Orgogozo 2008). However, as the data concerning the genetic sources of adaptive differences between species and populations accumulates this debate remains unresolved. Many examples exist of both structural and *cis*-regulatory changes driving adaptive differences (Hoekstra & Coyne 2007; Stern & Orgogozo 2008). Furthermore, as genome-wide surveys of sequence conservation have become possible, though few such surveys have been carried out, the evidence suggests that coding sequences show more adaptive changes than conserved non-coding elements (Halligan et al., 2013). Nevertheless, the results presented above are consistent with a greater role for regulatory change, at least in the evolution of craniofacial morphology.

An alternative explanation for the relatively low number of positively selection genes is that there is simply not enough power for these types of tests for positive selection. Power in these tests comes mainly from the number of lineages included and from an optimal divergence between species that results in enough phylogenetic signal within the sequences but does not saturate the number of synonymous substitutions (Kosiol et al., 2008). The number of species used here (maximum 8) is a moderate sample size in comparison to other studies (e.g. Kosiol et al., 2008; Foote et al., 2015). Additionally, manual inspection of some alignments and the estimates of *dS* in many cases suggest relatively low amounts of divergence between the species. A solution might be to expand the comparative approach and take advantage of other available passerine genomes to investigate clade specific molecular evolution within the crow lineage. This reduces the number of branches tested and increases power to detect differential molecular evolution among corvids.

Aside from the effects of poor alignments on inferences made from analyses of molecular evolution (Schneider et al., 2009), an important caveat is that these results are dependent on the tree topology. The procedure used here, concatenating alignments of a set of genes to produce a matrix from which a tree is estimated, is a standard method to estimate species trees (Kubatko & Degnan 2007; Mirarab et al., 2014; Roch & Steel 2015). However, this methods has been shown to often produce trees that are inconsistent with the true species tree (Kabutko & Degnan 2007; Roch & Steel 2015). Alternative methods have been developed which involve first estimating individual gene trees and summarising these into species trees (e.g. ASTRAL; Mirarab et al., 2014).

Additionally, the use of "core" genes, which are by definition highly conserved across many lineages, may be reducing the phylogenetic signal available in aligned sequences. Another approach might be to a set of well annotated and complete genes from throughout the genome and producing gene trees separately before summarising these into a species tree. Recent work has also shown that gene tree incongruence can have large effects on false positive rates in tests of differential substitution rates among lineages (Mendes & Hahn 2016).

The results of a gene family evolution analysis suggested many expansions and contractions. However, these results should be interpreted with caution. These analyses are heavily influenced by the quality and completeness of different genome anntoations. Although extensive curation and identification of orthologous relationships of the *C. cornix* annotation has been carried out (Poelstra et al., 2014), there remain many uncharacterised coding sequences and the *C. moneduloides* annotation is perhaps even less reliable. Nevertheless, concentrating on the more modest expansion and contraction events identifies some well described gene families such as the PAK (p21-activated kinases) family. PAK genes function, among other things, in immunity (Zhao & Manser 2012), a physiological function that is frequently seen to be under strong positive selection across many taxa (e.g. Sackton et al., 2007; Salazar-Jaramillo et al., 2014; Zhang et al., 2014). These results suggest that, although not all instances of expansions and contractions are false positives, it seems likely that gene families which show large differences in gene number reflect incomplete or poorly assembled gene families rather than extreme rates of gene gain and loss. Clearly, more accurate annotations as well as more *de novo* annotations from more species will be needed.

6.4.2 Signatures of Selection in Diversity

Genome-wide patterns of nucleotide diversity are assessed for the sister species *C. moneduloides* (the NC crow) and *C. woodfordi.* These species show differences in patterns of nucleotide diversity (π) and Tajima's *D.* Both species show levels of diversity in the same range as those reported for other passerine species (Balakrishnan & Edwards 2009; Huynh et al., 2010; Ellegren et al., 2012). The lowest levels of diversity are seen on the Z chromosome, along with the highest levels of F_{ST}, this is as expected given the differences in N_e between the Z chromosome and autosomes.

Additionally, the ratio of Z chromosomal to autosomal diversity is close to the 0.75 expected from theory (Corl & Ellegren 2012; Ellegren et al., 2013). In general, Tajima's *D* is higher in *C. woodfordi* than in the NC crow throughout the genome. Typically, values of Tajima's *D* above 0 are interpreted as a signal of balancing selection (Huber and Lohmueller 2016).

However, the fact that Tajima's *D* is elevated throughout the genome points to processes which alter diversity in the species as a whole rather than selection at specific loci. Possibilities include recent population expansions and contractions or population structure in the sample (Gattepaille et al., 2013; Huber and Lohmueller 2016). Alternatively, the patterns may indicate an issue with the sampling scheme and highlight the need for additional sampling throughout the geographic distribution of the species. For example, the NC crow samples constitute four individuals from the relatively isolated island of Maré and one individual from the main island Grand Terre. At the same time, previous work using microsatellite markers noted significant, fine-scale population differentiation among populations on Grand Terre (Rutz et al., 2012). Thus, population structure in the samples could conceivably produce artefacts in the data. Finally, of the *C. woodfordi* samples, three are from the island of Guadalcanal (one museum specimen caught in 1995) and the remaining two samples are from the island of Santa Isabel (both museum specimens from 1995). Thus in the *C. woodfordi* samples too, there is the potential for population structure to a be confounder in interpretation.

Potentially more reliable inference could be made using Approximate Bayesian Computation (ABC) methods like *fastsimcoal2* (Excoffier et al., 2013) or *δαδι* (Gutenkunst et al., 2009) and others (Csilléry et al., 2010) which allow the estimation of demographic parameters (like the effective population size) from the genomic data rather than relying on assumptions from the literature to parameterise simulations. This would give a more empirical neutral expectation of e.g. Tajima's *D* based on the available data. These methods can also test the hypothesis that the overall levels of, for example, Tajima's *D* are consistent with models of recent population contractions or expansions (Csilléry et al., 2010). However, the accuracy of these methods are likely to depend on the quality of the sampling of the species. In this case, the combination of few samples from relatively restricted parts of the range coupled with sampling non-

contemporary populations (i.e. museum specimens) is likely to introduce biases here as well.

In both species diversity is negatively correlated with chromosome length. This pattern is expected from theory that π should correlate with recombination rates, which are higher on microchromosomes (Ellegren 2007; Ellegren 2013). This pattern is seen in several other species of birds (Huynh et al., 2009; Aslam et al., 2012). Levels of diversity are also closely related to F_{ST} throughout the genome. This is also expected because F_{ST} is a relative measure of diversity. Similar patterns are seen in other crow systems as well (Vijay et al., 2016). Finally, π is highly correlated in orthologous windows throughout the genomes of the NC crow and *C. woodfordi*. This is not necessarily expected in relatively diverged species if the processes that shape genomic patterns of diversity become lineage specific after speciation and lineage sorting is complete.

However, correlations of diversity in orthologous genomic regions across species have been observed elsewhere. This is true in earlier (Vijay et al., 2016) and later (Dutoit et al., 2017) stages of the "speciation continuum." These patterns are attributed to the relatively stable karyotypes in avian lineages (Ellegren et al., 2013) and, potentially, to shared recombination landscapes (Vijay et al., 2016; Dutoit et al., 2017). However, in the absence of recombination maps for any of the species in this analysis this hypothesis is difficult to test. For the NC crow recombination maps may be difficult to generate owing to the intractability of keeping captive individuals in great numbers. However, the Hawai'ian crow population is kept entirely in captivity with a full pedigree of several hundred individuals, presenting an excellent opportunity to map recombination events throughout the Hawai'ian crow genome. Fine mapping of crossover events has recently been accmplished in a pedigree of only 11 flycatcher (*Ficedula albicollis*; Smeds et al., 2016) This type of population also lends itself more intuitively to GWAS style analysis of interesting traits (e.g. bill shapes).

Several regions show strong reductions in both Tajima's *D* and Fay and Wu's *H* within the NC crow and, crucially, not in *C. woodfordi*. These are taken as the regions showing the strongest evidence of recent selective sweeps. A survey of the region within 1Mb of the windows showing the largest reductions in Tajima's *D* finds many associated genes. Some (*Dkk2, Calm1*) are involved in the development of beak

morphologies in other birds (Wu et al., 2006; Abzhanov et al., 2006; Brugmann et al., 2010). Indeed, *Dkk2* shows the highest levels of differential expression in a comparison of chickens and ducks. Ducks also have unusually straight bills compared to other birds (Brugmann et al., 2010). Yet other genes in these regions are known to be expressed in the zebra finch brain. One of these genes, *Psen1*, has previously been associated with cognitive impairments, such as long-term memory potentiation, Alzheimer's and dementia, in both humans and mice (Morley & Montgomery 2001). Finally, *Foxo6* is a *forkhead box* transcription factor which is related to *Foxp2*. *Foxp2*, in humans, shows a strong signal of recent positive selection and an association with the ability to acquire speech (Enard et al., 2002). Similarly, in zebra finches (Teramitsu & White, 2006) and in great tits (Laine et al., 2016) there is an association with vocal learning and evidence for recent selective sweeps. Meanwhile, mutant forms of *Foxo6* have been associated with variation in the severity of schizophrenia symtpoms in humans (Shenker et al., 2017). Knockdown of expression in mice results in normal learning but impaired memory (Salih et al., 2012). In birds, the only reports are of elevated expression in the breast muscle in analyses of economically important traits of domestic ducks (Xu et al., 2012).

Finally, the observation that these regions are not associated with particularly striking peaks of F_{ST} is noteworthy. F_{ST} is generally high throughout the genome, which is unsurprising in a comparison of different species. Since there has likely been no gene flow between *C. woodfordi* and the NC crow since their divergence, this pattern is consistent with progressive homogenisation of F_{ST} throughout the genome as the "speciation continuum" progresses (Feder et al., 2012). In the *Heliconius* system comparisons of progressively more distantly related populations and species results in a more homogenous F_{ST} landscape throughout the genome with fewer obvious peaks (Nadeu et al., 2012; Martin et al., 2013). The same is true in a multi-species, multi-population study of crows in the *C. corone* species complex (Vijay et al., 2016). The species used in this chapter are 'good species' in that speciation is probably complete and gene flow rare or absent. Thus F_{ST} may not be the most appropriate statistic for identifying adaptively diverging regions *a priori.* Nevertheless these results bear on the debate surrounding the interpretation of peaks of differentiation (Noor & Bennett, 2009; Wolf & Ellegren 2016). In this case, searching for these peaks would be quite

uninformative because F_{ST} is rather uniform throughout the genome. At the same time, F_{ST} seems strongly related to π even in these already differentiated 'good species,' indicating a persistent role for factors other than gene flow and selection in producing the landscape of differentiation (Dutoit et al., 2017). Instead signatures of selection in reduced diversity or an excess of derived alleles appear more reliable to detect adaptively important differences. However, these approaches too are subject to some issues (e.g. sensitivity to demographic effects and variation in recombination or mutation rates throughout the genome) that may confound interpretation.

In addition to the problems highlighted above, it is difficult to associate the troughs of Tajima's *D* with any gene in particular that could be related to important phenotypes in the NC crow. Several genes occur in these regions. Nevertheless, the above genes represent good candidates where variants could be under selection to alter brain and neuron development or function, as well as craniofacial morphology. Follow up of these candidates is needed. Though transcriptome data are difficult to obtain for these species for practical and ethical reasons, especially for brain tissue, such data from the NC crow (especially during development as an embryo) would be invaluable. Alternatively, a GWAS style analysis could be performed in order to associate genome-wide markers with variation in morphological phenotypes in the NC or Hawai'ian crow populations. If loci strongly associated with traits of interest (e.g. beak shapes) localise to these same regions that would provide further evidence of adaptively important variants in these regions. Given the extensive knowledge we now have of the genes involved in determining beak and skull shapes, another approach might be to identify conserved regions up or downstream of these regions showing high levels of conservation. Such regions are likely to be regulatory domains and can be investigated for differences that correlate with beak shapes across avian lineages. This type of approach has been used successfully in studies on convergent evolution in rodent dentition (Tapaltsyan et al., 2016).

A final difficulty, unrelated to the choice of population genetic statistics, is the relatively poor knowledge of the genetic underpinnings of differences in cognitive traits (e.g. memory, learning, persistence) which are likely to be important in tool using behaviours (Emery & Clayton 2005; Kacelnik 2009; Holzhaider et al., 2010; Kenward et al., 2011). One difficulty arises even in defining these phenotypes in ways that make

them tractable for genomic analyses (de Geus et al., 2001; Sabb et al., 2009). Additionally, few studies have been published in non-human animal systems showing an association between gene variants and cognitive traits which makes *a priori* predictions difficult to formulate (but see Morley & Montgomery 2001). 'Reverse genetic' studies like the current one provide good ways of identifying candidate genes that can be followed up either in the same system or in model systems.

6.5 Concluding Remarks

In this chapter, I conduct a comparative genomic analysis of 15 species from the Corvid radiation with a particular focus on the tool-using New Caledonian crow (*C. moneduloides*). Rates of molecular evolution are found to be consistently elevated only in a handful of genes. Two of these are known to be associated with cognitive disorders and general measures of intelligence in humans and mice (*Htt* and *Dtnbp1*). Meanwhile, population genetic statistics from throughout the genome suggest that demographic forces may have played a substantial role in determining the patterns of genetic diversity. Tajima's *D* is somewhat elevated in *C. woodfordi* in comparison to the NC crow and in both species average Tajima's *D* on the autosomes is above zero. Nevertheless some signatures of selective sweeps are apparent even when using fairly stringent threshholds. These regions are also associated with strong reductions in Fay and Wu's *H* which is less sensitive to demographic forces. Several genes near these regions of selective sweeps are related to relevant phenotypes including beak and skull morphologies and cognition in mice, humans, and other songbirds. This study shows how a comparative approach can provide valuable insights into the loci contributing to important adaptive differences between species. It also highlights how contrasts between species can advance our understanding of the processes that shape genetic diversity throughout the genome in different demographic scenarios.

Chapter 7 Discussion and Conclusions

7.1 The Genomics of Adaptation

With the "modern synthesis" (Huxley 1942) of Darwinian natural selection and genetics came the understanding that changes in allele frequencies in populations are the mediators of evolutionary change. Although genetic investigations of populations have been underway since the early 19th century the last decade has seen a rapid proliferation of technology which has allowed the investigation of the entire genome of organisms. Thus, the comparative method has been extended to comparative genomics (Ellegren 2008; Pardo-Diaz et al., 2015). Although there have been promising results in the identification of loci that are important in species or population differences there remain many challenges in unerstanding the ways in which different processes contribute to patterns of variation seen in natural (Noor & Bennet 2009; Wolf & Ellegren 2016; Ravinet et al., *in press*) and experimental populations (Kofler & Schlötterer 2014; Baldwin-Brown et al., 2014; Kessner & Novembre 2015).

There has been some criticism of the whole enterprise of attempts to identify the genomic loci underlying adaptively important traits (Rockman 2012; Travisano & Shaw 2013). The criticism is mainly that loci which contribute to variation in traits, even adaptive traits, within QTL mapping populations are numerous and of small individual effect. Many studies have indeed shown that this is the case (Mackay et al., 2009; Rockman 2012). Therefore, it has been argued, attempts to identify and account for the molecular basis of adaptive traits are in vain and in any case would not greatly advance our understanding of the general principles of evolution (Rockman 2012; Travisano & Shaw 2013). However, while QTLs explaining variation within populations may be elucsive, the loci that matter for differences between populations may in fact have large effects and be discoverable (Rausher & Delph 2015). Since it is the variants that underlie differences between populations and species which are the "stuff of evolution" (Rausher & Delph 2015), methods that directly address population differences should have a greater chance of identifying loci that are important in population and species

divergence. With the ever decreasing costs of whole-genome sequencing it is now possible to identify variants segregating in multiple populations or species which differ in evolutionarily important traits (e.g. Jones et al., 2012; The *Heliconius* Genome Consortium 2012; Poelstra et al., 2014; Lamichhaney et al., 2015; Vijay et al., 2016; Božičević et al., 2016). Identifying consistent differences between independent comparisons of populations, or signatures of selection within one population can identify loci which are important in producing differences (e.g. Martins et al., 2014; Vijay et al., 2016).

At the same time there is more to the genetics of adaptation than simply the identification of "the genes" involved. While the question of "why" selection is acting on phenotypes is important and insight is unlikely to come from a genetic perspective, the question of "how" selection acts equally merits investigation. How does selection act to build genomes? What roles do regions of reduced recombination (e.g. inversions) have in spread of beneficial co-adapted alleles (Hoffmann & Riesberg 2008)? How does selection act to mediate sexual conflict and produce sexually dimorphic traits (Wilkinson et al., 2015)? These are questions to which genomic studies can provide valuable insights, in some cases without identifying variants (genes/alleles, rearrangements, or single nucleotides) contributing to phenotypic differences.

In this thesis I have undertaken a comparative genomic approach to investigate the genomic differences between populations in various systems. I take advantage of different species, and both natural and laboratory populations. Below, I outline the main findings, synthesise some common conclusions and discuss opportunities for follow-up work.

7.2 Summary of Findings

In chapter 2 I considered the challenge of identifying consistent allele frequency changes in response to selection in experimental evolution studies. Experimental evolution studies have been a popular tool to study adaptation to novel environments for many years. With the rise of low-cost next-generation sequencing (NGS) technologies the prospects for understanding the genomics of adaptive change have become greatly improved. Therefore, there is a need to develop both reliable experimental protocols and

analytical approaches using these technologies that can answer these questions. A critical feature of experimental evolution studies is to use replicated experimental treatments. This replication allows researchers to distinguish between changes that may simply be due to chance (genetic drift) and consistent changes across multiple replicates which are more likely to have been driven by parallel selection (Kawecki et al., 2012). In genomic terms this amounts to identifying consistent allele frequency changes across replicate lines. Several methods have been developed to test for such consistent change in allele frequencies (Kofler et al., 2011; Baldwin-Brown et al., 2014), but a consensus has not been reached on the best approach (Baldwin-Brown et al., 2014; Kessner & Novembre 2015).

I conducted a population genetic simulation to test different methods which have been proposed for the analysis of experimental evolution genome data and identified some serious problems with popular approaches. I also proposed alternative approaches which performed well under the null hypothesis with no cost in the power to detect true positives. This chapter highlights the importance of evaluating the performance of new analytical approaches by simulation. In genomics this can be particularly powerful because the population genetic theory of the behaviour of allele frequencies under the null hypothesis is well established (Charlesworth & Charlesworth 2008).

In chapter 3 I analysed genome sequencing data from an ongoing, long-term experimental evolution study in *D. pseudoobscura* (Crudgington et al 2005). 8 replicate lines were set up in 2002 to investigate the response to selection under altered mating systems. Four lines were assigned to an "elevated polyandry" (E) treatment and four to a "enforced monogamy" (M) treatment (see chapter three for more details). The aim was to identify genomic loci (SNPs) which showed a consistent allele frequency difference between the E and M treatments across replicates. To this end, I applied the results of my simulation study (Chapter 2) and performed a Generalised Linear Model (GLM) with quasibinomial error distribution for each SNP with experimental treatment as a fixed effect.

The results suggest that consistent allele frequency differences localised to clusters of highly differentiated SNPs. Additionally, there was an excess of differentiated SNPs on the 3rd chromosome and the X chromosome arms. Meanwhile, population genetic summary statistics (Tajima's *D,* F_{ST}) suggest that these regions are

not obviously associated with "peaks" of F_{ST} but that they show clear signatures of selective sweeps in one or the other treatment lines. This highlights the importance of using a range of measures to make inferences about within population processes (such as selection and recombination variation) and contrast them with between population measures of differentiation (Wolf & Ellegren 2016). Candidate SNPs are associated with many genes which have mutant phenotypes that are relevant to previous phenotypic assays of these experimental evolution lines (e.g. courtship song, male aggression, and male post-copulatory manipulation of females). Additionally, patterns of diversity and differentiation on the X chromosome and autosomes suggest that evolution has been faster on the X. Differentiation (F_{ST}) is generally higher on the X than on the autosomes in E and M lines but the difference is greater in E lines. Additionally, X:A ratio of π is lower in E lines. These results are consistent with a faster-X (due to a greater efficiency of selection) and an effect of polyandry on reducing the ratio of Ne but they are not consistent with a strong effect of sexual selection or conflict which should increase the ratio of diversity (Ellegren 2009; Vicoso & Charlesworth 2009; Corl & Ellegren 2012).

In chapter 4 I considered a novel approach to identifying associations between genomic markers and a phenotype. This involved using isofemale lines that vary in a phenotype of interest, namely the re-mating rate. SNPs that are consistently fixed for alternative alleles in isofemale lines that are at opposite extremes of the phenotypic distribution should include loci which are linked to causal loci. In any pairwise comparison many fixed differences will occur by chance but if several pairwise comparisons are performed using pairs of lines from different source populations the rate of spurious associations should be reduced. I used population genetic simulations to establish how many consistently fixed differences are expected between any pairs of lines sampled from an ancestral population. I compared the observed number of consistently fixed differences to the simulated distributions and concluded that there are more such differences than expected by chance. Many of these fixed differences also occur near or in genes which have previously been associated with female re-mating rates in *Drosophila* species. Although sample sizes were extremely small, and results should be treated with caution this methods seems a fruitful approach to identifying genotype-phenotype associations.

In chapter 5 I took advantage of several populations of *D. montana* from North America and Finland. I used Principle Component (PC) analysis to determine the main axes of climatic variation across these populations. I then relate differentiation at SNPs to environmental differentiation across these populations. The aim was to uncover SNPs and nearby genes which show a relationship with climatic variation to uncover candidate genes that may be important in the divergence of these populations. Many SNPs show a significant relationship with climatic variation and those that do are associated with genes involved in the circadian rhythm and also in cell membrane integrity and cryoprotection. Also, many of the genes near these candidate SNPs lie in regions which show characteristic signals of selective sweeps (reduced Tajima's *D*) indicating a role for local adaptation. In this chapter too I apply quasibinomial GLMs (form chapter 2) but with a continuous predictor of allele frequencies. However, distributions of p-values suggest that such a linear model is a poor fit to the data. This could be because there are many non-linear relationships between the continuous predictors and allele frequencies which are not captured by a linear model. At the same time there are likely many linked loci which result in p-values at nearby SNPs being correlated. Bayesian methods such as BayeScEnv (de Villemereiul & Gaggiotti 2015) seemed to perform much better in these analyses.

Finally, in chapter 6 I switch focus to a comparative study of crows. The New Caledonian (NC) crow is a tropical island endemic that has become the focus of much research due to its tool-using behaviour. In this study I undertake a comparative genomic study using whole genome sequences from 15 species of crows. The use of multiple crow species with similar ecological niches (tropical island endemics) allows the distinction of signatures of adaptation associated with island colonisation from those associated with the particular characteristics of the NC crow. The results suggest limited support for elevated rates of coding sequence evolution within the NC crow lineage. However, a few genes associated with cognitive disease (schizophrenia and Huntington's disease) in humans and mice show a robust signal of positive selection. At the same time, multiple regions throughout the genome show reductions of Tajima's *D* and Fay and Wu's *H* which are greater than expected from neutral population genetic simulations. These regions lie near genes involved in the development of craniofacial morphology in birds and mice as well as other cognition related genes. However, there

is evidence of population genetic structure in these samples, probably owing to the sampling of birds from different islands so results should be treated with caution. At the same time these data show that in a comparison of closely related but "good" species (i.e. speciation is complete) there is a strong correlation in the level of diversity within orthologous windows. This highlights a potentially strong effect of forces other than selection in driving patterns of differentiation and diversity even on longer evolutionary timescales.

7.3 General Discussion and Future Work

7.3.1 Recombination, Inversions and Hitchhiking

The results from the above chapters also bear on the expected effect of recombination and hitchhiking on genomic patterns of diversity and differentiation. For example, the observation in two chapters (Chapter 3 and Chapter 5) that candidate SNPs seem highly clustered raises the possibility of adaptive colocalisation in low recombination regions or inversions, though genetic hitchhiking effects cannot at present be ruled out. In particular, chromosomal inversions are increasingly being recognised as important aspects of genomic organisation that can have an impact on adaptive evolution (Kirkpatrick & Barton, 2006; Hoffman & Riesberg 2008). For example, adaptive roles for inversions have been observed in natural population clines of *Drosophila melanogaster* (Kapun et al., 2016), seaweed flies (*Coelopa frigida*; Wellenreuther et al., 2017); monkey flowers (*Mimulus guttatus*; Oneal et al., 2014), and others. In this context, future work could investigate the presence and locations of inversions and recombination hot/cold-spots among the experimental evolution lines (Chapter 3) and wild populations of *D. montana* (Chapter 5).

In *Drosophila* systems it should be feasible to investigate the presence of inversions and conduct breeding experiments to quantify recombination throughout the genome. In *D. pseudoobscura* several inversions are known to segregate within the species (Sturtevant & Dobzhansky 1936; Dobzhansky & Sturtevant 1937) while others are fixed between *D. pseudoobscura* and the sister species *D. perisimils* (Noor et al., 2001; Stevison et al., 2011). Stevison et al., (2011) found that although gene flow between species is possible (via double recombination) through fixed inversions it is

likely to be exceedingly rare. Between species inversions have maintained high divergence within these inversions on geological timescales (McGaugh & Noor 2012). Meanwhile these inversions are associated with various traits conferring a degree of reproductive isolation between the species (Noor et al., 2001), pointing to a large role for inversions in speciation. In *D. montana*, reproductive isolation exists between populations and is apparently mediated, at least in part, by cuticular hydrocarbons and courtship song (Jennings et al., 2011; 2014). QTLs for differences in song characteristics and genital morphology between populations localise to chromosomal regions known to be polymorphic for inversions (Morales-Hojas et al., 2007; Schäfer et al., 2011; Lagisz et al., 2012). These data point to a role for inversions in maintaining population differences in this species but more precise characterisation of inversions and their distributions are needed. Such data would provide further insight into the processes that drive patterns of differentiation and allele frequency change seen in these systems.

The observation in Chapter 6 of a close correlation in patterns of diversity within orthologous windows in already "good" species (i.e. speciation is complete) is unexpected. This result highlights the importance of knowing about recombination landscapes in drawing inferences about patterns of diversity and differentiation form genomic data (Dutoit et al., 2017). If landscapes are shared this could result in correlated patterns and must be taken into account when drawing inferences about the action of other evolutionary forces (Dutoit et al., 2017). Landscapes are known to be shared across species in the zebra finch and close relatives (Singhal et al., 2015). Measuring recombination is difficult in wild populations of New Caledonian crows which are difficult to keep and breed in captivity. However, the captive population of Hawai'ian crows is more amenable on account of the larger sample sizes and known pedigree. Therefore, comparison of the recombination landscape with the landscape of diversity within Hawai'ian crows would shed light on this issue. For example, recombination patterns have been well characterised in flycatchers (Kawakami et al., 2014; Smeds et al., 2016), and in zebra finches (Singhal et al., 2015). Comparison of these data to that from the Hawai'ian crow would help to determine whether this landscape is conserved across species and contributes to intraspecific variation (Dutoit et al., 2017).

7.3.2 Functional Genomics: Knockdown, Knockout, and Knockin

While "knock-out" mutations (loss of function) to alleles have contributed much to our understanding of which genes affect phenotypes they offer rather crude functional characterisation and many knock-out individuals don't survive long enough for phenotypes to be assessed. In contrast, "knock-down" treatments (e.g. RNAi) allow a more fine-grained estimation of the effects of different levels of expression of a gene. If differences in expression are seen between two treatments then expression can be artificially altered to test if the phenotype responds as predicted. Finally genome editing, or "knock-in," allows the replacement of one genetic variant for another to test whether the phenotype responds as predicted. The recently introduced technology of CRISPR/*Cas9* promises to ease the creation of knock-out, knock-down, and knock-in lines (Bono et al., 2015).

Many of the genes identified as being in close linkage with candidate SNPs seem to have associated mutant phenotypes that are relevant to expected biological functions, such as sexual selection and conflict (Chapter 3 and Chapter 4), circadian rhythms and the control of diapause or cold tolerance (Chapter 5), and craniofacial morphologies as well as cognitive traits (Chapter 6). These genes therefore represent exciting candidates for further functional validation with knock-out/-down/-in techniques like CRISPR/*cas9* or RNAi. Indeed in the *D. montana* system such experimental approaches have already been performed for the gene *Inos* which is located in a region that shows signs of ongoing selection in high altitude and high latitude populations (Chapter 4). *Inos*, which codes for the protein *myo*-inositol, is observed at high concentrations in overwintering *D. montana*, and is associated with cold-tolerance in other insects (Vesala et al., 2012). Knockdown of *Inos* by RNAi in *D. montana* does not reduce the ability for cold acclimation but significantly increases cold-specific rates of mortality. Individuals acclimatised at 5°C die at higher rates than individuals acclimatised at 19°C (Vigoder et al., 2016). Other candidates might include the gene *vri* which is involved in circadian rhythms among *Drosophila* and regulates the negative feedback loop of another circadian rhythm gene, *Clock* (*clk*) (Rosato et al., 2006; Helfrich-Förster 2017), shows differential expression in response to cold acclimation in *D. montana* (Vesala et al.,

2012c; Parker et al., 2015), and also occurs in a region showing signs of selective sweeps (Chapter 4).

These types of experiments have their limitations however. For example, the creation of transgenic lines is much more practical and ethical in *Drosophila* than in birds, let alone endangered species like the Hawai'ian crow. Even when it is in principle possible, the creation of transgenic lines is laborious so testing large numbers of genes is impractical, especially for whole-organism phenotypes. These methods are still mostly available only for models for which protocols are well established (i.e. *Drosophila melanogaster;* Bassett & Liu 2014), but it seems likely that they will soon be more widely available also for non-model species such as *D. montana* and *D. pseudoobscura*.

Additionally, these types of functional genomic experiments need a precise target. For example. altering an entire gene by introducing a copy with particular amino acid substitutions or by altering precise parts of the regulatory region. This type of fine scale modification is important because subtle changes can underlie important adaptive differences. For example, in inter-species QTL mapping of courtship song characters (presumably important aspects of assortative mating and reproductive isolation), the gene *fruitless* is frequently implicated (Gleason & Ritchie 2004; Lagisz et al., 2012). At the same time, positive selection at this locus is restricted to alternatively spliced exons implying that alternative splicing and specific isoforms of genes may underlie adaptive differences between species (Parker et al., 2013). Such differences would be obscured if a locus was simply knocked out altogether, or expression in general is reduced in a non-isoform-specific manner. Similarly, a classic case of regulatory change underlying morphological differences is the regulation of the gene *shavenbaby* (*svb*), a transcript of the *ovo* locus, which is involved in the development of trichomes on *Drosophila* larvae (Stern & Frankel 2013). Expression of *svb* is controlled from a series of enhancers up to ~90-140kb from the start of transcription (McGregor et al., 2007; Frankel et al., 2012). Parallel changes in two enhancers explain much of the convergent loss of trichomes in the comparison of *D. melanogaster* and *D. sechellia,* as well as *D. littoralis* and *D. ezoana* (Frankel et al., 2012).

Finally, although an altered locus produces a measurable phenotypic change, this may not necessarily imply that it is an important locus for e.g. reproductive isolation, or

sexual conflict. This phenotypic change must further translate to a difference in assortative mating or coercive/manipulative ability. Such differences will only be measured in mate choice or social interaction experiments, or fitness assays in ecologically relevant environments with wild-type and mutant individuals. With genome editing such experiments are becoming feasible (Bono et al., 2015). The power of this approach can be seen in previous work in model systems. McGregor et al., (2007) created transgenic *D. melanogaster* by replacing the enhancers of *svb* (see above) with those of *D. sechellia.* Although the regions replaced were quite wide, the resulting transgenic *D. melanogaster* larvae expressed a phenotype similar to that of *D. sechellia* confirming the locations of variants underlying species differences in the enhancer regions (McGregor et al., 2007). Similarly, in African populations of *D. melanogaster* there is adaptive variation in cuticular melanism which is partly controlled by the *ebony* locus (Rebeiz et al., 2009). In an *ebony* null mutant background, transgenes from "light" and "dark" lines were able to rescue a range of abdomen shades (Rebeiz et al., 2009). These studies highlight the power of transgenics in quantifying the effect of specific sequence variants on adaptive differences between species and populations.

In this context, the observation that regions around *Inos* look to be under selection in northern populations of *D. montana,* similar to patterns first noticed around *ebony* (Rebeiz et al., 2009), would predict that variation in sequence that alters transcription is important for population differences. Thus, if differences are identified and a southern variant can be introduced to otherwise northern genomic background, individuals with a southern version of this gene should fare worse. This would constitute a good demonstration that variants in or near this gene which alter transcription levels during cold acclimation are functionally important for population divergence in cold-adaptation traits. Similar assays could be performed for other loci which show signs of selection in different climates (Chapter 5), and also differential expression during cold acclimation or diapause. Similarly, knockout or replacement of genes involved in the formation of wings or wing muscle in *D. pseudoobscura* could be replaced in E and M female backgrounds (Chapter 3) to test for an effect on e.g. song characters.

7.3.3 Islands of Differentiation

Finally, the results in this thesis bear on the debate surrounding the expected patterns of diversity and differentiation observed in diversifying wild populations (Noor & Bennet 2009; Wolf & Ellegren 2017; Ravinet et al., *in press*). Peaks of F_{ST} are often interpreted as a combination of divergent selection at loci underlying adaptation or reproductive isolation and gene flow in the rest of the genome (Noor & Bennett 2009; Wolf & Ellegren 2017). However, a debate is still ongoing over the interpretation of these peaks of differentiation. Such peaks can arise due to divergent selection and gene flow but also, for example, as a product of segregating ancestral variation and variation in recombination throughout the genome, in the absence of gene flow (Noor & Bennet 2009).

In chapter 3 there was no gene flow across treatment lines to homogenise the F_{ST} landscape. Though there were some small peaks of F_{ST} between pairs of treatment lines these were not useful diagnostic features for regions with significant allele frequency differences between treatments. While these regions showed obvious effects in consistent allele frequency differences between E and M lines, F_{ST} was not obviously greater than elsewhere in the genome. Thus localised differentiation (allele frequency differences) due to selection can build up without the homogenising effects of gene flow elsewhere in the genome. These regions also showed very clear signs of selective sweeps within the treatment lines (reduced Tajima's D and π). Higher pairwise F_{ST} was not a useful diagnostic feature of these regions. These results are consistent with a model where peaks of divergence often seen between populations or hybridising species are the result of divergent selection acting on a few loci and homogenising gene flow. One explanation for a fairly homogenous landscape of F_{ST} is that the relatively short timescales (compared to evolutionary history) may be too short in experimental studies. Even ~200 generations may not be enough time to result in large peaks of F_{ST}, especially when stochastic forces (drift) are strong due to small population sizes within treatment lines driving elevation of F_{ST} throughout the genome even though selection is only acting in specific regions.

Similar results are clear from chapter 5. Regions in which F_{ST} was associated with variation in climate across populations often showed signatures of positive selection, at least in some populations. Thus, a range of summary statistics can tell us

more about the processes (such as selection and hitchhiking within populations) shaping patterns of diversity than a single statistic (e.g. F_{ST}), unfortunately these are rarely used (Wolf & Ellegren 2017). Unfortunately, the debate remains unsolved mainly because of a lack of knowledge regarding the recombination landscape, a crucial part of alternative hypotheses of islands of differentiation. However, these experimental evolution lines provide an excellent system in which to study the forces that drive differentiation in localised regions.

7.4 Concluding Remarks

Understanding the genetic basis of phenotypic differences between populations/species has been constrained by our ability to identify the loci underlying phenotypic variation within species (Rockman 2012; Travisano & Shaw 2013). By contrast comparative and population genomics can provide high resolution lanscapes of polymorphism within the genome and facilitate contrasts between populations/species. Such landscapes can directly identify the signatures of selection and demographic forces acting on regions of the genome to identify the loci underlying population/species differences. In this thesis I have taken a comparative approach to investigate the genomics of adaptation in several systems. I have also contributed novel approaches to the analysis of studies of stratified populations and experimental evolution studies. Overall, my results highlight the importance of a thorough survey of patterns of genetic variation when drawing inferences about the processes that have shaped them. These studies have also identified promising avenues for follow up studies. It is becoming increasingly clear that complete insights will ultimately come not from single studies that apply only methods such as the ones I use in this thesis, but from the accumulation of several independent studies combining different approaches (e.g. comparative genomics, functional genomics, phenotypic fitness assays, etc.). This work sheds light on the process of evolution at the genome scale in the specific systems studied here and contributes toward our understanding of population divergence, adaptation, and speciation in general.

References

Abzhanov, A. & Tabin, C.J. 2004. *Shh* and *Fgf8* act synergistically to drive cartilage outgrowth during cranial development. *Developmental Biology*. **273**: 134-148.

Abzhanov, A., Protas, M., Grant, B.R., Grant, P.R. & Tabin, C.J. 2004. *Bmp4* and morphological variation of beaks in Darwin's finches. *Science*. **305**: 1462-1465.

Adrion, J.R., Hahn, M.W. & Cooper, B.S. 2015. Revisiting classic clines in *Drosophila* in the age of genomics. *Trends in Genetics*. **31**: 434 – 444.

Abzhanov, A., Kuo, W.P., Hartmann, C., Grant, B.R., Grant, P.R. & Tabin, C.J. 2006. The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature*. **442**: 563-567.

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. Wiley-Interscience, New York.

Andersson, M. 1994. *Sexual Selection.* Princeton University Press, Princeton, New Jersey.

Andrews, S., 2014. FastQC: A quality control tool for high throughput sequence data. [Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/].

Anholt, R.R.H., Lyman, R.F. & Mackay, T.F.C. 1996. Effects of single P-Element insertions on olfactory behavior in Drosophila melanogaster. *Genetics* **143**: 293–301.

Ansorge, W. 2009. Next-generation DNA sequencing techniques. *New Biotechnology*. **25**: 195 – 203.

Arabiza, L., Gottipati, S., Siepel, A. & Keinan, A. 2014. Contrasting X-linked and autosomal diversity across 14 human populations. *American Journal of Human Genetics*. **94**: 827-844.

Arif, S., Hilbrant, M., Hopfen, C., Almudi, I., Nunes, M.D.S., Posnien, N., Kuncheria, L., Tanaka, K., Mitteroecker, P., Schlötterer, C. & Mcgregor, A.P. 2013. Genetic and developmental analysis of differences in eye and face morphology between *Drosophila simulans* and *Drosophila mauritiana*. *Evolution & Development*. **15**: 257-267.

Arnqvist, G. & Nilsson, T. 2000. The evolution of polyandry: multiple mating and female fitness in insects. *Animal Behaviour*. **60**: 145-164.

Arnqvist, G. and Rowe, L. 2005. *Sexual Conflict*. Princeton University Press, Princeton, New Jersey.

Arquier, N., Bourois, M., Colomban, J. & Léopold, P. 2005. *Drosophila* Lk6 kinase controls phosphorylation of eukaryotic translation initiation factor 4E and promotes normal growth and development. *Current Biology*. **15**: 19–23.

Aspi, J. & Hoikkala, A. 1995. Male mating success and survival in the field with respect to size and courtship song characters in *Drosophila littoralis* and *D. montana* (Dipter: Drosophilidae). *Journal of Insect Behavior*. **8**: 67–87.

Avila, F.W., Cohen, A.B., Ameerudeen, F.S., Duneau, D., Suresh, S., Mattei, A.L. & Wolfner, M.F. 2015. Retention of ejaculate by *Drosophila melanogaster* females requires the male-derived mating plug protein PEBme. *Genetics*. **200**: 1171 – 1179.

Ayllon, F., Kjærner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M.F., Dahle, G., Taranger, G.L., Glover, K.A., Almén, M.S., Rubin, C.J., Edvardsen, R.B., & Wargelius, A. 2015. The vgll3 Locus Controls Age at Maturity in Wild and Domesticated Atlantic Salmon (*Salmo salar* L.) Males. *PLoS Genetics,* **11**: 1–15.

Bacigalupe, L.D., Crudgington, H.S., Hunter, F., Moore, a. J. & Snook, R.R. 2007. Sexual conflict does not drive reproductive isolation in experimental populations of Drosophila pseudoobscura. *Journal of Evolutionary Biology* **20**: 1763–1771.

Baena-López, L.A., Baonza, A. & García-Bellido, A. 2005. The orientation of cell divisions determines the shape of *Drosophila* organs. *Current Biology*. **15**: 1640 – 1644.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. & Noble, W.S. 2009. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research* **37**: 202–208.

Bakanidze, G., Brandl, E.J., Hutzler, C. Aurass, F., Onken, S., Rapp, M.A. & Puls, I. 2016. Association of Dystrobrevin-binding protein 1 polymorphisms with sustained attention and set-shifting in Scizophrenia patients. *Neuropsychobiology*. **74**: 41-47.

Balakrishnan, C.N. & Edwards, S.V. 2009. Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the Zebra Finch (*Taniopygia guttata*). *Genetics*. **181**: 645-660.

Balakrishnan, C.N., Edwards, S.V. & Clayton, D.F. 2010. The Zebra Finch genome and avian genomics in the wild. *Emu*. **110**: 233-241.

Balding, D. 2003. Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology,* **63**: 221–230.

Baldwin-Brown, J.G., Long, A.D. & Thornton K.R. 2014. The power to detect quantitative trait loci using resequenced experimentally evolved populations of diploid, sexual organisms. *Molecular Biology and Evolution*. **31**: 1040–1055.

Bansal, V., 2010. A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics.* **26**: 318–324.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P., Marth, G.T. 2011. BamTools: a C++ API and toolkit for analysing and managing BAM files. *Bioinformatics.* **27**: 1691–1692.

Barrett, R.D.H., Rogers, S.M. & Schluter, D. 2008. Natural selection on a major armor gene in threespine Stickleback. *Science*. **322**: 255-257.

Bassett, A.R., Tibbit, C., Ponting, C. & Liu, J-L. 2013. Highly efficient targeted mutagenesis of *Drosophila* with the CRISPR/*Cas9* system. *Cell Reports*. **4**: 220 – 228.

Bassett, A.R. & Liu, J-L. 2014. CRISPR/Cas9 and genome editing in *Drosophila*. *Journal of Genetics and Genomics*. **41**: 7-19.

Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stöbe, P., Futschik, A.*, &* Schlötterer, C. .2013. A Genome-Wide, Fine-Scale Map of Natural Pigmentation Variation in *Drosophila melanogaster*. *PLoS Genetics,* **9**: e1003534.

Beckenbach, A.T. 1981. Multiple mating and the "Sex-Ratio" trait in *Drosophila pseudoobscura*. *Evolution*. **35**: 275-281.

Beckenbach, A.T. 1996. Selection and the "Sex-Ratio" polymorphism in natural populations of *Drosophila pseudoobscura*. *Evolution*. **50**: 787-794.

Ben-Shahar, Y. 2011. Sensory functions for degenerin/epithelial sodium ion channels (DEG/ENaC). *Advances in Genetics.* **76**: 1-26

Bergland, A.O., Behrman, E.L., O'Brien, K.R., Schmidt, P.S., Petrov, D.A. 2014. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila. PLoS Genetics.* **10**: e1004775.

Bergland, A., Tobler, R., Gonzáles, J., Schmidt, P. & Petrov, D. 2016. Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. *Molecular Ecology.* **25**: 1157 – 1174.

Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M. & Roth, F.P. 2009. Next generation software for functional trend analysis. *Bioinformatics* **25**: 3043–3044.

Blanchardon, E., Grima, B., Klarsfeld, A., Chélot, E., Hardin, P.E. & Préat, T. 2001. Defining the role of *Drosophila* lateral neurons in the control of circadian activity and eclosion rhythms by targeted genetic ablation and PERIOD protein overexpression. *European Journal of Neuroscience*. **13**: 878-888

Boake, C.R.B., Arnold, S.J., Breden, F., Meffert, L.M., Ritchie, M.G., Taylor, B.J., Wolf, J.B. & Moore, A.J. 20502. Genetic tools for studying adaptation and the evolution of behavior. *The American Naturalist*. **160**: S143-159.

Boitard, S., Schlo, C., Nolte, V., Pandey, R.V. & Futschik, A. 2012. Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples Research article. *Molecular Biology and Evolution,* **29**: 2177–2186.

Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*. **30**: 2114–2120.

Boll, W. & Noll, M. 2002. The *Drosophila Pox neuro* gene: control of male courtship behaviour and fertility as revealed by a complete dissection of all enhancers. *Development*. **129**: 5667–5681.

Bono, J.M., Olesnicky, E.C. & Matzkin, L.M. 2015. Connecting genotypes, phenotypes and fitness: harnessing the power of CRISPR/Cas9 genome editing. *Molecular Ecology*. **24**: 3810-3822.

Borge, T., Webster, M.T., Andersson, G. & Sætre G-P. 2005. Contrasting patterns of polymorphism and divergence on the Z chromsome and autosomes in two *Ficedula* flycatcher species. *Genetics*. **171**: 1861-1873.

Boulton, R.A. & Shuker, D.M. 2013. Polyandry. *Current Biology*. **23**: 1080-1081.

Bousquet, F., Nojima, T., Houot, B., Chauvel, I., Chaudy, S. & Dupas, S. 2012. Expression of a desaturase gene, *dsat1*, in neural and nonneural tissues separately affects perception and emission of sex pheromones in *Drosophila*. *PNAS*. **109**: 249–254.

Boutros, M., Kiger, A.A., Armknecht, S., Kerr, K., Hild, M., Koch, B., Haas, S.A., Heidelberg Fly Array Consortium, Paro, R. & Perrimon, R. 2004. Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science*. **303**: 832 – 835.

Božičević, V., Hutter, S., Stephan, W. & Wollstein, A. 2016. Population genetic evidence for cold adaptation in European *Drosophila melanogaster* populations. *Molecular Ecology*. **25**: 1175-1191.

Bradbury, D., Smithson, A. & Krauss, S.L. 2013. Signatures of diversifying selection at EST-SSR loci and association with climate in natural *Eucalyptus* populations. *Molecular Ecology*. **22**: 5112–5129.

Broad Institute. *no date.* Picard Tools [Available from: http://broadinstitute.github.io/picard/].

Brugmann, S.A., Powder, K.E., Young, N.M., Goodnough, L.H., Hahn, S.M., James, A.W., Helms, J.A. & Lovett, M. 2010. Comparative gene expression analysis of avian embryonic facial structures reveals new candidates for human craniofacial disorders. *Human Molecular Genetics*. **19**: 920-930.

Burdfield-Steele, E.R. & Shuker, D.M. 2014. Mate-guarding in a promiscuous insect: species descrimination influences context-dependent behaviour. **28**: 1031–1042.

Butlin, R., Debelle, A., Kerth, C., Snook, R.R., Beukeboom, L.W., Castillo Cajas, R.F., Diao, W., Maan, M.E., Paolucci, S., Weissing, F.J., van de Zande, L., Hoikkala, A., Geuverink, E., Jennings, J., Kankare, M., Knott, K.E., Tyukmaeva, V.I., Zoumadakis, C., Ritchie, M.G., Barker, D., Immonen, E., Kirkpatrick, M., Noor, M., Macias Garcia, C., Schmitt, T. & Schilthuizen, M. 2012. What do we need to know about speciation. *Trends in Ecology and Evolution*. **27**: 27–39.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics*. **10**: 421.

Campesan, S., Dubrova, Y., Hall, J.C. & Kyriacou, C., 2001. The nonA gene in *Drosophila* conveys species-specific behavioral characteristics. *Genetics.* **158**: 1535–1543.

Carneiro, M., Baird, S.J.E., Afonso, S., Ramirez, E., Tarroso, P., Teotónio, H., Villafuerte, R., Nachman, M.W. & Ferrand, N. 2013. Steep clines within a highly permeable genome across a hybrid zone between two subspecies of the European rabbit. *Molecular Ecology*. **22**: 2511–2525.

Carreira, V.P., Soto, I.M., Mensch, J. & Fanara, J.J. 2011. Genetic basis of wing morphogenesis in *Drosophila*: sexual dimorphism and non-allometric effects of

shape variation. *BMC Developmental Biology*. **11**: 32

Carroll, S.B. 2005. Evolution at two levels: on genes and form. *PLoS Biology*. **3**: e245.

Castresana, J. 2000. Selection of conserved blocks from multiple alignment for their use in phylogenetic analysis. *Molecular Biology and Evolution*. **17**: 540-552.

Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal Jr., G.., Shapiro, M.D., Brady, S.D., Southwick, A.M., Absher, D.M., Grimwood, J., Schmutz, J., Myers, R.M., Petrov, D., Jónsson, B., Schluter, D., Bell, M.A. & Kingsley, D.M. 2010. Adaptive evolution of pelvic reduction of a *Pitx1* enhancer. *Science*. **327**: 302–306.

Chapman, T., Arnqvist, G., Bangham, J. & Rowe, L. 2003. Sexual conflict. *Trends in Ecology and Evolution*. **18**: 41-47.

Chapman, T. 2008. The soup in my fly: Evolution, form and function of seminal fluid proteins. *PLoS Biology*. **6**: e176.

Charlesworth, B. & Charlesworth, D. 2008. Elements of Evolutionary Genetics. Robers and Compnay, Greenwood Village, Colorado.

Cheah, S-Y., Lawford, B.R., Young, R.M., Morris, C.P. & Voisey, J. 2015. Dysbindin (DTNBP1) variants are associated with hallucinations in scizophrenia. *European Psychiatry*. **30**: 486-491.

Chen, J., Källman, T., Ma, X., Gyllenstrand, N., Zaina, G., Morgante, M., Bousquet, J., Eckert, A., Wegrzyn, J., Neale, D., Lagercrantz, U. & Lascoux, M. 2012. Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce (*Picea abies*). *Genetics*. **191**: 865–881.

Chen, J., Källman, T., Ma, X-F., Zaina, G., Morgante, M. & Lascoux, M. 2016. Identifying genetic signatures of natural selection using pooled population sequencing in *Picea abies*. *G3*. **6**: 1979-1989.

Chen, Y., Lee, S.F., Blanc, E., Reuter, C., Werheim, B., Marinez-Diaz, P., Hoffmann, A.A. & Partridge, L. 2012. Genome-wide transcription analysis of clinal genetic variation in *Drosophila*. *PLoS ONE*. **7**: e34620.

Cheng, C., White, B., Kamdem, C., Mockaitis, K., Constantini, C., Hahn, M.W. & Besansky, N.J. 2012. Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population re-sequencing approach. *Genetics*. **190**: 1417–1432.

Cichon, S., Mühleisen, T.W., Degenhardt, F.A., Mattheisen, M., Miró, X., Strohmaier,

J., Steffens, M., Meesters, C., Herms, S., Weingarten, M., Priebe, L., Haenisch, B., Alexander, M., Vollmer, J., Breuer, R., Schmäl, C., Tessmann, P., Moebus, S., Wichmann, H.E., Schreiber, S., Müller-Myhsok, B., Lucae, S., Jamain, S., Leboyer, M., Bellivier, F., Etain, B., Henry, C., Kahn, J.P., Heath, S., Hamshere, M., O'Donovan, M.C., Owen, M.J., Craddock, N., Schwarz, M., Vedder, H., Kammerer-Ciernioch, J., Reif, A., Sasse, J., Bauer, M., Hautzinger, M., Wright, A., Mitchell, P.B., Schofield, P.R., Montgomery, G.W., Medland, S.E., Gordon, S.D., Martin, N.G., Gustafsson, O., Andreassen, O., Djurovic, S., Sigurdsson, E., Steinberg, S., Stefansson, H., Stefansson, K., Kapur-Pojskic, L., Oruc, L., Rivas, F., Mayoral, F., Chuchalin, A., Babadjanova, G., Tiganov, A.S., Pantelejeva, G., Abramova, L.I., Grigoroiu-Serbanescu, M., Diaconu, C.C., Czerski, P.M., Hauser, J., Zimmer, A., Lathrop, M., Schulze, T.G., Wienker, T.F., Schumacher, J., Maier, W., Propping, P., Rietschel, M. & Nöthen, M. 2011. Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *American Journal of Human Genetics,* **88**: 372–381.

Clayton, N. & Emery, N.J. 2005. Corvid cognition. *Current Biology*. **15**: 80-81.

Clough, E., Jimenez, E., Kim, Y., Whitworth, C., Neville, M.C.., Hempel, L.U., Pavlou, H.J., Chen, Z., Sturgill, D., Dale, R.K., Smith, H.E., Przytycka, T.M., Goodwin, S.F., van Doren, M. & Oliver, B. 2014. Sex- and Tissue-Specific Functions of Drosophila Doublesex Transcription Factor Target Genes. *Developmental Cell* **31**: 761–773.

Cnotka, J. Güntürkün, O., Rehkämper, G., Gray, R.D. & Hunt, G.R. 2008. Extraordinary large brains in tool-using New Caledonian crows (*Corvus moneduloides*). *Neuroscience Letters*. **433**: 241-245.

Cochran, W.G. 1954. Some Methods for Strengthening the Common $\chi 2$ Tests. *Biometrics,* **10**: 417–451.

Colinet, H., Lee, S.F. & Hoffman, A. 2010. Temporal expression of heat shock genes during cold stress and recovery from chill coma in adult *Drosophila melanogaster. The FEBS Journal.* **277**: 174-185

Colosimo, P.F., Peichel, C.L., Nereng, K., Blackman, B.K., Shapiro, M.D., Schluter, D. & Kingsley, D.M. 2004. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biology*. **2**: e0635.

Colosimo, P.F., Hoseman, K.E., Balabhadra, S., Villarreal Jr., G., Dickson, M.,

Grimwood, J., Schmutz, J., Myers, R.M., Schluter, D. & Kingsley, D. 2005. Widespread parallele evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*. **307**: 1928-1933.

Corl, A. & Ellegren, H. 2012. The genomic signature of sexual selection in the genetic diversity of sex chromosomes and autosomes. *Evolution*. **66**: 2138-2149.

Cortesi, F., Musilová, Z., Stieb, S.M., Hart, N.S., Siebeck, U.E., Malmstrøm, M., Tørreson, O.K., Jentoft, S., Cheney, K.L., Marshall, N.J., Carlton, K.L. & Salzburger, W. 2015. Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes. *PNAS*. **112**: 1493-1498.

Cosetti, M., Culang, D., Kotla, S., Brien, P.O. & Daniel, F. 2008. Unique transgenic animal model for hereditary hearing loss. *Annals of Otology, Rhinology, and Laryngology*. **117**: 827–833.

Crawley M.J. 2013. The R Book (2nd Edition). John Wiley & Sons Ltd, Chichester, West Sussex, PO19 8SQ, United Kingdom.

Crudgington, H., Beckerman, A.P., Brüstle, L., Green, K., & Snook, R.R. 2005. Experimental removal and elevation of sexual selection: Does sexual selection generate manipulative males and resistant females? *American Naturalist.* **165**: S72 – S87.

Crudgington, H.S., Fellows, S. & Snook, R.R. 2010. Increased opportunity for sexual conflict promotes harmful males with elevated courtship frequencies. *Journal of Evolutionary Biology* **23**: 440–446.

Crudgington, H.S., Fellows, S., Badcock, N.S. & Snook, R.R. 2009. Experimental manipulation of sexual selection promotes greater male mating capacity but does not alter sperm investment. *Evolution* **63**: 926–938.

Cruickshank, T.E. & Hahn, M.W. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*. **23**: 3133-3157.

Csilléry, K., Blum, M.G.B., Gaggiotti, O.E. & François, O. 2010. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*. **25**: 410 – 418.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L.,

Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M.J., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R. & Flicek, P. 2015. Ensembl 2015. *Nucleic Acids Research*. **43**: D662–D669.

Cunningham, C.B., Badgett, M.J., Meagher, R.B., Orlando, R. & Moore, A.J. 2017. Ethological principles predict the neuropeptides co-opted to influence parenting. *Nature Communications*. **8**: 30-36.

Dabney, A. & Storey, J.D. 2015. qvalue: Q-value estimation for false discovery rate control. R package. Available from: http://github.com/jdstorey/qvalue.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R., Lunter, G., Marth, G., Sherry, S.T., McVean, G., Durbin, R. & 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics*. **27**: 2156–2158.

Darwin, C. 1859 *On the Origin of Species, or, the Preservation of Favoured Races in the Struggle for Life.* John Murray, London.

Darwin, C. 1871 *The Descent of Man and Selection in Relation to Sex*. John Murray, London

David, J.R., Gibert, P., Legout, H., Pétavy, G., Capy, P. & Moreteau, B.2005. Isofemale lines in *Drosophila*: an empirical approach to quantitative trait analysis in natural populations. *Heredity*. **94**: 3-12.

de Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*. **22**: 1269–1271.

de Geus, E.J.C., Wright, M.J., Martin, N.G. & Boomsma, D.I. 2001. Genetics of brain function and cognition. *Behavior Genetics*. **31**: 489-495.

de Villemereuil, P. & Gaggiotti, O.E. 2015. A new F_{ST}-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*. **6**: 1248 – 1258.

Debelle A. 2013. *The role of sexual selection in the evolution of reproductive isolation.* Ph.D. Thesis, University of Sheffield, Sheffield, U.K.

Debelle, A., Ritchie, M.G. & Snook R.R. 2016. Sexual selection and assortative mating: an experimental test. *Journal of Evolutionary Biology*. **29**: 1307–1316.

Degoutin, J.L., Milton, C.C., Yu, E., Tipping, M., Bosveld, F., Yang, L., Bellaiche, Y., Veraksa, A. & Harvey, K.F. 2013. Riquiqui and minibrain are regulators of the hippo pathway downstream of dachsous. *Nature Cell Biology*. **15**: 1176–1185.

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A.., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D. & Daly, M.J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**: 491–498.

Dey, B.K., Zhao, X., Popo-Ola, E. & Campos, A.R. 2009. Mutual regulation of the *Drosophila* disconnected (*disco*) and Distal-less (*Dll*) genes contributes to proximal-distal patterning of antenna and leg. *Cell Tissue Research*. **338**: 227– 240.

Dobzhansky, T.H, & Sturtevant, A.H. 1937. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics*. **23**: 28–64.

dos Santos, G., Schroeder, A.J., Goodman, J.L., Strelets, V.B., Crosby, M. A., Thurmond, J., Emmert, D.B. & Gelbart, W.M. 2014. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research,* **43**: D690–D697.

Dubnau, J., Chiang, A-S., Grady, L., Barditch, J., Gossweiler, S., McNeil, J., Smith, P., Buldoc, F., Scott, R., Certa, U., Broger, C. & Tully, T. 2003. The *staufen/pumilio* pathway is involved in *Drosophila* long-term memory. *Current Biology*. **13**: 286-296.

Dutoit, L., Vijay, N., Mugal, C.F., Bossu, C.M., Burri, R., Wolf, J.B.W. & Ellegren, H. 2017. Covariation in levels of nucleotide diversity in homologous regions of the avian genome long after completion of lineage sorting. *Proceedings of the Royal Society of London: Biology*. **284**: 20162756.

Ebacher, D.J.S., Todi, S.V., Eberl, D.F. & Falk, G.E.B. 2007. *cut* mutant *Drosophila* auditory organs differentiate abnormally and degenerate. *Fly*. **1**: 86–94.

Edwards, A.C., Zwarts, L., Yamamoto, A., Callaerts, P. & Mackay, T.F.C. 2009. Mutations in many genes affect aggressive behaviour in *Drosophila melanogaster*. *BMC Biology.* **7**: 29.

Ekseth, O.K., Kuiper, M. & Mironov, V. 2014. orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics*. **30**: 734–736.

Ellegren, H. 2007. The molecular evolutionary genomics of birds. *Cytogenetics and Genome Research*. **117**: 120-130.

Ellegren, H. 2008. Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*. **17**: 4586-1596.

Ellegren, H. 2009. The different levels of genetic diversity in sex chromosomes and autosomes. *Trends in Genetics*. **25**: 278-284.

Ellegren, H. 2013. The evolutionary genomics of birds. *Annual Review of Ecology, Evolution and Systematics*. **44**: 239-259.

Ellegren, H. 2014 Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*. **29**: 51-63.

Ellegren, H., Smeds, L., Burri, R., Olason, P.I., Backström, N., Kawakami, T., Künstner, A., Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., Uebbing, S. & Wolf, J.B.W. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*. **491**: 756-760.

Emery, N.J. & Clayton, N. 2004. The mentality of crows: convergent evolution of intelligence in corvids and apes. *Science*. **306**: 1903-1907.

Emlen, D.J. & Nijhout, H.F. 2000. The development and evolution of exaggerated morphologies in insects. *Annual Review of Entomology*. **45**: 661–708.

Emlen, D.J. 2008. The evolutoin of animal weapons. *Annual Review of Ecology, Evolution, and Systematics*. **39**: 387–413.

Emms, D.M. & Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, Genome Biology **16**:157.

Enard, W. Lai, Cecilia S.L., Wiebe, V., Kitano, T., Monaco, A.P. & Pääbo, S. 2002. Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature*. **418**: 869-872.

Endler, J.A. 1973. Gene flow and differentiation. *Science*. **179**: 243–250.

Endler, J.A. 1977. *Geographic Variation, Speciation, and Clines.* Princeton University Press, Princeton, New Jersey.

Enikolopov, G. Banerji, J. & Kuzin, B. 1999. Nitric oxide dependent *Drosophila*

development. *Cell Death and Differentiation*. **6**: 956–963.

Enright, A., Dongen, S. & Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*. **30**: 1575–1584.

Evans, J.P. & Simmons, L.W. 2008. The genetic basis of traits regulating sperm competition and polyandry: can selection favour the evolution of good- and sexy-sperm? *Genetica*. **134**: 5–19.

Evans, T.A., Santiago, C., Arbeile, E. & Barshaw, G.J. 2015. *Robo2* acts in *trans* to inhibit *Slit-Robo1* repulsion in pre-crossing comissural axons. *eLife*. **4**: e08407.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C. & Foll, M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genetics*. **9**: e1003905.

Fabian, D.K., Kapun, M., Nolte, V., Kofler, R., Schmidt, P.S., Schlötterer, C. & Flatt, T. 2016. Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology*. **21**: 4748 – 4769.

Feder, A.F., Petrov, D. & Bergland, A. 2012. LDx: Estimation of linkage disequilibrium from high throughput pooled resequencing data. *PLoS ONE*. **7**: e48588.

Feder, A.F., Petrov, D.A. & Bergland, A.O. 2012. LDx: Estimation of linkage disequilibrium from high-throughput pooled resequencing data. *PLoS ONE*. **7**: e48588.

Feder, J.L., Egan, S.P. & Nosil, P. 2012. The genomics of speciation-with-gene-flow. *Trends in Genetics*. **28**: 342-350.

Ferguson, L., Lee, S.F., Chamberlain, N., Nadeau, N., Joron, M., Baxter, S., Wilkinson, P., Papanicolaou, A., Kumar, S., Kee, T.J., Clark, R., Davidson, C., Glithero, R., Beasley, H., Vogel, H., Ffrench-Constant, R., Jiggins, C. 2010. Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the *HmYb/Sb* locus. *Molecular Ecology*. **19**: 240 – 254.

Ferretti, L., Ramos-Onsins, S.E. & Pérez-Enciso, M. 2013. Population genomics from pool sequencing. *Molecular Ecology,* **22**: 5561 – 5576.

Fitzpatrick, M.J., Ben-Shahar, Y., Smid, H.M., Vet, L.E.M., Robinson, G.E. & Sokolowski, M.B. 2005. Candidate genes for behavioural ecology. *Trends in Ecology and Evolution*. **20**: 96 – 104.

Flatt, T. 2016. Genomics of clinal variation in *Drosophila*: Disentangling the interactions of selection and demography. *Molecular Ecology*. **25**: 1023 – 1026.

Foll, M. & Gaggiotti, O.E. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics*. **180**: 977 – 993.

Foll, M., Poh, Y-P., Renzette, N., Ferrer-Admetlla, A., Bank, C., Shim, H., Malaspinas, A-S., Ewing, G., Liu, P., Wegmann, D., Caffrey, D.R., Zeldovich, K.B., Bolon, D.N., Wang, J.P., Kowalik, T.F., Schiffer, C.A., Finberg, R.W. & Jensen, J.D. 2014. Influenza drug virus resistance: a time-sampled population genetics perspective. *PLoS Genetics*. **10**: e1004185.

Foote, A.D., Liu, Y., Thomas, G.W.C., Vinař, T., Alföldi, J., Deng, J., Dugan, S., van Elk, C.E., Hunter, M.E., Joshi, V., Khan, Z., Kovar, C., Lee, S.L., Lindblad-Toh, K., Mancia, A., Nielsen, R., Qin, X., Qu, J., Raney, B.J., Vijay, N., Wolf, J.B.W., Hahn, M.W., Muzny, D.M., Worley, K.C., Gilbert, M.T.P. & Gibbs, R.A. 2015. Convergent evolution of the genomes of marine mammals. **47**: 272-275.

Frankel, N., Wang, S. & Stern, D. 2012. Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. *PNAS*. **109**: 20975-20979.

Franssen, S.U., Nolte, V., Tobler, R. & Schlotterer, C. 2014. Patterns of Linkage Disequilibrium and Long Range Hitchhiking in Evolving Experimental *Drosophila melanogaster* Populations. *Molecular Biology and Evolution,* **32**: 495–509.

Gaertner, B.E., Ruedi, E.A., Mccoy, L.J., Moore, J.M., Wolfner, M.F. & Mackay, T.F.C. 2015. Heritable variation in courtship patterns in *Drosophila melanogaster*. *G3*. **5**: 531-539.

Gardiner, A., Barker, D., Butlin, R.K., Jordan, W.C. & Ritchie, M.G. 2008. *Drosophila* chemoreceptor gene evolution: selection, specialization and genome size. *Molecular Ecology*. **17**: 1648-1657.

Gerber, A.P., Luschnig, S., Krasnow, M.A., Brown, P.O. & Herschlag, D. 2006. Genome-wide identification of mRNAs associated with the translational regulator *pumilio* in *Drosophila melanogaster*. *PNAS*. **103**: 4487-4492.

Gerrard, D.T., Fricke, C., Edward, D.A., Edwards, D.R. & Chapman, T. 2013. Genome-wide responses of female fruit flies subjected to divergent mating regimes. *PLoS ONE*. **8**: e68136.

Giardina, T.J. 2015. Mating rate and the influence of female genetics on remating in

Drosophila melanogaster. PhD Thesis, Binghamton University.

Giardina, T.J., Beavis, A., Clark, A.G. & Fiumera, A.C. 2011. Female influence on pre- and pos-copulatory sexual selection and its genetic basis in *Drosophila melanogaster*. *Molecular Ecology*. **20**: 4098–4108.

Gibbs, A.G. 2002. Lipid melting points and cuticular permeability: new insights into an old problem. *Journal of Insect Physiology*. **48**: 391–400.

Giniger, E., Tietje, K., Jan, L.Y. & Jan, Y.N. 1994. *lola* encodes a putative transcription factor required for axon growth and guidance in *Drosophila*. *Development*: **120**: 1385-1398.

Gleason, J.M. & Ritchie, M.G. 2004. Do quantitative trait loci (QTL) for a courtship song difference between *Drosophila simulans* and *D. sechellia* coincide with candidate genes and intraspecific QTL? *Genetics*. **166**: 1303-1311.

Godenschwege, T. a., Reisch, D., Diegelmann, S., Eberle, K., Funk, N., Heisenberg, M., *et al.* 2004. Flies lacking all synapsins are unexpectedly healthy but are impaired in complex behaviour. *Eur. J. Neurosci.* **20**: 611–622.

Godenschwege, T.A., Reisch, D., Diegelmann, S., Eberle, K., Funk, N., Heisenberg, M., Hoppe, V., Hoppe, J., Klagges, B.R.E., Martin, J.R., Nikitina, E.A., Putz, G., Reifegerste, R., Reisch, N., Rister, J., Schaupp, M., Scholz, H., Schwärzel, M., Werner, U., Zars, T.D., Buchner, S. & Buchner, E. 2004. Flies lacking all synapsins are unexpectedly healthy but are impaired in complex behaviour. *European Journal of Neuroscience*. **20**: 611–622.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. & Pääbo, S. 2010. A draft sequence of the Neanderthal genome. *Science*. **328**: 710 – 722.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. 2009.

Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*. **5**: e1000695.

Guttman, M. & Rinn, J.L. 2012. Modular regulatory principles of large non-coding RNAs. *Nature*. **482**: 339–346.

Guven-Ozkan, T. Busto, G.U., Schutte, S.S., Cervantes-Sandoval, I., O'Dowd, D.K. & Davis, R.L. 2016. *Mir-980* is a memory suppressor MicroRNA that regulates the autism-susceptibility gene *A2bp1*. Cell Reports. **14**: 1698–1709.

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D.L., Houle, D., Charlesworth, B. & Keightley, P.D. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. **445**: 82-85.

Haerty, W., Jagadeeshan, S., Kulathinal, R.J., Wong, A., Ravi Ram, K., Sirot, L.K., Levesque, L., Artieri, C.G., Wolfner, M.F., Civetta, A. & Singh, R.S. 2007. Evolution in the fast lane: Rapidly evolving sex-related genes in *Drosophila*. *Genetics*. **177**: 1321-1335.

Hahn, M.W., de Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*. **15**: 1153–1160.

Haldane, J.B.S. 1948. The theory of a cline. *Journal of Genetics*. **48**: 277–284.

Halligan D.L., Kousathanas, A., Ness, R.W., Harr, B., Eöry, L., Keane, T.M., Adams, David J. & Keightley, P.D. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genetics.* **9**: e1003995.

Häring, E., Däubl, B., Pinsker, W., Kryukov, A. & Gamauf, A. 2012. Genetic divergences and intraspecific variation in corvids of the genus *Corvus* (Aves: Corvidae) – a first survey based on museum specimens. *Journal of Zooligical Systematics and Evolutionary Research*. **50**: 230-246.

Harshman, L.G., Hoffmann, A.A. & Prout, T.1988. Environmental effects on remating in *Drosophila melanogaster*. *Evolution*. **42**: 312-321.

Harvey, P.H. & Pagel, M.D. 1991. *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.

Harvey, P.H. & Purvis, A. 1991. Comparative methods for explaining adaptations. *Nature*. **351**: 619 – 624.

Hedges, S., Dudley, J. & Kumar, S. 2006. TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics*. **22**: 2971–2972.

Hekmat-Scafe, D., Scafe, C.R., McKinney, A.J. & Tanouye, M.A. 2002. Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. *Genome Research* **12**: 1357–1369**.**

Hellfrich-Förster, C., Edwards, T., Yasuyama, K., Wisotzki, B., Schneuwly, S., Stanewsky, R., Meinertzhagen, I.A. & Hofbauer, A. 2002. The extraretinal eyelet of *Drosophila*: Development, ultrastructure, and putative circadian function. *The Journal of Neuroscience*. **22**: 9255–9266.

Helfrich-Förster, C. 2017. The *Drosophila* clock system. Pp 133- in *Biological Timekeeping: Clocks, Rhythms, and Behaviour.* (V. Kumar Ed). Springer Nature, India.

Herrera, P., Taylor, M.L., Skeats, A., Price, T.A.R. & Wedell, N. 2014. Can patterns of inversions in *Drosophila pseudoobscura* predict polyandry across a geographical cline? *Ecology and Evolution.* **4**: 3072-3081.

Hijmans, R.E., van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J.A., Lamigueiro, O.P., Bevan, A., Racine, E.B. & Shortridge, A. 2016. raster: Geographic Data Analysis and Modeling. [Available from: https://CRAN.R-project.org/package=raster]

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology.* **25**:1965–1978.

Hoban, S., Kelley, J.L, Lotterhos, K.E., Antolin, M., F., Bradburd, G., Lowry, D.B., Poss, M.L., Reed, L.K., Storfer, A., Whitlock, M.C. 2016. Finding the genomic basis local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*. **188**: 379-397.

Hoekstra, H., Drumm, K. & Nachman, M. 2004. Ecological genetics of adaptive color polymorphism in pocket mice: geographic variation in selected and neutral genes. *Evolution*. **58**: 1329–1341.

Hoekstra, H.E., Hirschmann, R.J. Bundey, R.A., Insel, P.A. & Crossland, J.P. 2006. A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*. **313**: 101-104.

Hoekstra, H.E. & Coyne, J.A. 2007. The locus of evolution: Evo-devo and the genetics of adaptation. *Evolution*. **61**: 995-1016.

Hoffman, A. & Riesberg, L. 2008. Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology, Evolution and Systematics*. **39**: 21–42.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A. & Cresko, W.A. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetic.* **6**: e1000862.

Hoikkala, A., Klappert, K. & Mazzi, D. 2005. Factors affecting male song evolution in *Drosophila montana*. *Current Topics in Developmental Biology*. **67**: 225–250.

Holland, B. & Rice, W.R. 1998. Chase-away sexual selection: antagonistic seduction versus resistance. *Evolution*. **52**: 1–7.

Holland, B. & Rice, W.R. 1999. Experimental removal of sexual selection reverses intersexual coevolution and removes a reproductive load. *PNAS*. **96**: 5083 – 5088.

Hollis, B., Houle, D., Yan, Z., Kawecki, T.J. & Keller, L. 2012. Evolution under monogamy feminizes gene expression in *Drosophila melanogaster*. *Nature Communications*. **5**: 3482**.**

Hollis, B., Houle, D. & Kawecki, T.J. 2016a. Evolution of reduced post-copulatory molecular interactions in *Drosophila* populations lacking sperm competition. *Journal of Evolutionary Biology.* **29**: 77-85.

Hollis, B., Keller, L. & Kawecki, T.J. 2016b. Sexual selection shapes development and maturation rates in *Drosophila*. *Evolution*. **71**: 304-314.

Homyk, T. & Sheppard, D.E. 1977. Behavioral mutants of *Drosophila melanogaster*. I. Isolation and mapping of mutations which decrease flight ability. *Genetics*. **87**: 95–104.

Hosken, D.J. & Ward, P.I. 2001. Experimental evidence for testis size evolution via sperm competition. *Ecology Letters*. **4**: 10–13.

Hosken, D.J., Garner, T.W.J., Ward, P.I. 2001. Sexual conflict selects for male and female reproductive characters. *Current Biology*. **11**: 489 – 493.

Hu, N. & Castelli-Gair, J. 1999. Study of the posterior spiracles of *Drosophila* as a model to understand the genetic and cellular mechanisms controlling morphogenesis. *Developmental Biology*. **210**: 197–201.

Huang Y., Wright S.I. & Agrawal, A.F. 2014. Genome-wide patterns of genetic variation within and among alternative selective regimes. *PloS Genetics,* **10**: e1004527.

Huang, D.A.W., Sherman, B.T. & Lempicki, R.A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**: 44–57.

Huang, D.A.W., Sherman, B.T. & Lempicki, R.A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene sets. *Nucleic Acids Research*. **37**: 1–13.

Huang, D.W., Sherman, B.T. & Lempicki, R.A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*. **4**: 44–57.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A., Turlapati, L., Zichner, T., Zhu, D., Lyman, R.F., Magwire, M.M., Blankenburg, K., Carbone, M.A., Chang, K., Ellis, L.L., Fernandez, S., Han, Y., Highnam, G., Hjelmen, C.E., Jack, J.R., Javaid, M., Jayaseelan, J., Kalra, D., Lee, S., Lewis, L., Munidasa, M., Ongeri, F., Patel, S., Perales, L., Perez, A., Pu, L-L., Rollmann, S.M., Ruth, R., Saada, N., Warner, C., Williams, A., Wu, Y-Q., Yamamoto, A., Zhang, Y., Zhu, Y., Anholt, R.R.H., Korbel, J.O., Mittelman, D., Muzny, D.M., Gibbs, R.A., Barbadilla, A., Johnston, J.S., Stone, E.A., Richards, S., Deplancke, B. & Mackay, T.F. C. 2014. Natural varation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Research.* **24**: 1193-1208.

Huber, C.D. & Lohmueller, K.E. 2016. Population genetic tests of neutral evolution. Pp 112 – 118 in *The Encyclopedia of Evolutionary Biology Vol. 3* (Eds. R.M. Kilman). Oliver Walter.

Hudson, R.R., Kreitman, M. & Aguadé, M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics*. **116**: 153 – 159.

Hunt, G.R. & Gray, R.D. 2003. Diversification and cumulative evolution in New Caledonian crow tool manufacture. *Proceedings of the Royal Society: B.* **270**: 867-874.

Huxley J. 1938. Clines: an auxiliary taxonomic principle. *Nature*. 142: 219–220.

Huxley, J.S. 1942. *Evolution: The Modern Synthesis.* 3rd Edition. George Allen &

Unwin Ltd., London.

Huynh, L.Y., Maney, D.L. & Thomas, J.W. 2010. Contrasting population genetic patterns within the white-throated sparrow genome (*Zonotrichia albicollis*). *BMC Genetics*. **11**: 96

Immonen, E. & Ritchie, M.G. 2012. The genomic response to courtship song stimulation in female *Drosophila melanogaster*. *Proceedings of the Royal Society: Biological Sciences*. **279**: 1359-1365.

Immonen, E., Snook, R.R. & Ritchie, M.G. 2014. Mating system variation drives rapid evolution of the female transcriptome in Drosophila pseudoobscura. *Ecology and Evolution* **4**: 2186–2201.

Immonen, E., Sayadi, A., Bayram, H. & Arnqvist, G. 2017. Mating changes sexually dimorphic gene expression in the seed beetle *Callosobruchus maculatus*. *Genome Biology and Evolution*. **9**: 677-699.

Innocenti, P. & Morrow, H. 2009. Immunogenic males: a genome-wide analysis of reproduction and the cost of mating in *Drosophila melanogaster* females. *Journal of Evolutionary Biology*. **22**: 964-973.

Iranmehr, A., Akbari, A., Shlötterer, C. & Bafna, V. 2016. CLEAR: composition of likelihoods for evolve and resequence experiments. *Genetics*. **XX**:XXXX–XXXXX.

Ivanov, D.K., Escott-Price, V., Ziehm, M., Magwire, M.M., Mackay, T.F.C., Partridge, L. & Thornton, J.M. 2015. Longevity GWAS using the *Drosophila* Genetic Reference Panel. *Journals of Gerontology: Biological Sciences*. **70**: 1470-1478.

Janicke, T., Sandner, P., Ramm, S.A., Vizoso, D.B. & Schärer, L. 2016. Experimentally evolved and phenotypically plastic responses to enforced monogamy in a hermaphroditic flatworm. *Journal of Evolutionary Biology*. **29**: 1713 – 1727.

Jennings J.H., Mazzi, D., Ritchie, M.G. & Hoikkala, A. 2011. Sexual and postmating reproductive isolation between allopatric *Drosophila montana* populations suggest speciation potential. *BMC Evolutionary Biology*. **11**: 68.

Jennings, J.H., Snook, R.R. & Hoikkala, A. 2014. Reproductive isolation among allopatric *Drosophila montana* populations. *Evolution*. **11**: 3095-3108.

Jensen, J.D. & Bachtrog, D. 2011. Characterising the influence of effective population size on the rate of adaptation: Gillespie's Darwin Domain. *Genome Biology and Evolution*. **3**: 687-701

Jha, A.R., Zhou, D., Brown, C.D., Kreitman, M., Haddad, G.G. & White, K.P. 2016. Shared genetic signals of hypoxia adaptation in *Drosophila* and high altitude human populations. *Molecular Biology and Evolution*. **33**: 501–517.

Jiggins, C.D., Wallbank, R.W.R. & Hanly, J.J. 2017. Waiting in the wings: what can we learn about gene co-option from the diversification of butterfly wing patterns? *Philosophical Transactions of the Royal Society: Biology*. **372**: 20150485.

Jiggins, F.M. & Kim, K-W. 2006. Contrasting evolutionary patterns in *Drosophila* immune receptors. *Journal of Molecular Evolution*. **63**: 769-780

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. & Charpentier, E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*. **337**: 816 – 822.

Jónás Á., Taus, T., Kosiol, C., Schlötterer, C. & Futschik, A. 2016. Estimating the effective population size from temporal allele frequency changes in experimental evolution. *Genetics*. **204**: 723-735.

Jones, F., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen, A.A., Howes, T., Amemiya, C., Baldwin, J., Bloom, T., Jaffe, D.B., Nicol, R., Wilkinson, J., Lander, E.S., di Palma, F., Lindblad-Toh, K. & Kingsley, D.M. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*. **484**: 55-61.

Joron, M., Frezal, L., Jones, R.T., Chamberlain, N.L., Lee, S.F., Haag, C.R., Whibley, A., Becuwe, M., Baxter, S.W., Ferguson, L., Wilkinson, P.A., Salazar, C., Davidson, C., Clark, R., Quail, M.A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., Jones, M.C., Rogers, J., Jiggins, C.D., Richard, H. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*. **477**: 203 – 206.

Joron, M., Papa, R., Beltrán, M., Chamberlain, N., Mavárez, J., Baxter, S., Abanto, M., Bermingham, E., Humphray, S.J., Rogers, J., Beasley, H., Barlow, K., ffrench-Constant, R.H., Mallet, J., McMillan, W.O., Jiggins, C.D. 2006. A conserved supergene locus controls colour pattern diversity in *Heliconius* butterfly. *PLoS Biology*. **4**: e303.

Jønsson, K.A., Fabre, P.H. & Irestedt, M. 2012. Brains, tools, innovation and biogeography in crows and ravens. *BMC Evolutionary Biology*. **12**: 72.

Kabutko, L.S. & Degnan, J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*. **56**: 17-24.

Kacelnik, A. 2009 Tools for thought or thoughts for tools? *PNAS*. **106**: 10071-10072.

Kahle, D. & Wickham H. ggmap: Spatial Visualization with ggplot2. *The R Journal*, **5**:144–161.

Kamath, R.S. & Ahringer, J. 2003. Genome-wide RNAi screening in *Caeonrhabditis elegans*. *Methods*. **30**: 313 – 321.

Kang, L., Aggarwal, D.D., Rashkovetsky, E., Korol, A.B. & Michalak, P. 2016. Rapid genomic changes in *Drosophila melanogaster* adapting to dessiccation stress in an experimental evolution system. *BMC Genomics,* **17**: 233.

Kankare, M., Salminen, T., Laiho, A., Vesala, L. & Hoikkala, A., 2010. Changes in gene expression linked with adult reproductive diapause in a northern malt fly species: a candidate gene microarray study. *BMC Ecology*. **10**: 3.

Kankare, M., Parker, D.J., Merisalo, M., Salminen, T. & Hoikkala, A. 2016. Transcriptional differences between diapausing and non-diapausing *D. montana* females reared under the same photoperiod and temperature. *PLoS ONE.* **11**: e0161852.

Kapun, M., Van Schalkwyk, H., McAllister, B., Flatt, T. & Schlötterer, C. 2014. Inference of chromosomal inversion dynamics from Pool-Seq data in natural and laboratory populations of Drosophila melanogaster. *Molecular Ecology,* **23**: 1813–1827.

Kapun, M., Fabian, D.K., Goudet, J. & Flatt, T. 2016. Genomic evidence for adaptive inversion clines in *Drosophila melanogaster*. **33**: 1317-1336.

Kaspar, M., Schneider, M., Chia, W. & Klein, T. 2008. Klumpfuss is involved in the determination of sensory organ precursors in *Drosophila. Developmental Biology*. **324**: 177–191.

Kauranen, H., Ala-Honola, O., Kankare, M. & Hoikkala, A. 2016. Circadian clock of *Drosophila montana* is adapted to high variation in summer day lengths and temperatures prevailing at high latitudes. *Journal of Insect Physiology*. **89**: 9–18.

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C.F., Olason, P. & Ellegren, H. 2014. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology*. **23**: 4035-4058.

Kawecki, T., Lenski, R.E., Ebert, D., Hollis, B., Olivieri, I. & Whitlock M.C. 2012. Experimental evolution. *Trends in Ecology and Evolution*. **27**: 547–560.

Keightley, P.D., Ness, R.W., Halligan, D.L., & Haddrill, P.D. 2014. Estimation of the spontaneous mutation rate in a *Drosophila melanogaster* full-sib family. *Genetics,* **196**: 313–320.

Keightley, P.D., Pinharanda, A., Ness, R.W., Simpson, F., Dasmahapatra, K.K., Mallet, J., Davey, J.W. & Jiggins, C.D. 2015. Estimation of the spontaneous mutaiton rate in *Heliconius melpomene*. *Molecular Biology and Evolution*. **32**: 239-243.

Kenward, B., Weir, A.A.S., Rutz, C. & Kacelnik, A. 2005. Tool manufacture by naïve juvenile crows. *Nature*. **433**: 121.

Kenward, B., Rutz, C., Weir, A.A.S. & Kacelnik, A. 2006. Development of tool use in New Caledonian crows: inherited action patterns and social influences. *Animal Behaviour.* **72**: 1329-1343.

Kessner, D. & Novembre, J. 2015. Power analysis of artificial selection experiments using efficient whole genome simulation of quantitative traits. *Genetics*. **199**: 991-1005.

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

Kirkpatrick, M. & Barton, N. 1997. Evolution of a species' range. *The American Naturalist.* **150**: 1–23.

Kirkpatrick, M. & Barton, N. 2006. Chromosome inversions, local adaptation and speciation. *Genetics*. **173**: 419-434.

Klappert, K., Mazzi, D., Hoikkala, A. & Ritchie, M.G. 2007. Male courtship song and female preference variation between phylogenetically distinct populations of *Drosophila montana*. *Evolution*. **61**: 1481–1488.

Kleinjan, D.A. & van Heyningen, V. 2005. Long-range control of gene expression: Emergin mechanisms and disruption in disease. *American Journal of Human*

Genetics. **76**: 8-32.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C., Schlötterer, C. 2011. PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE.* **6**: e15925.

Kofler, R., Pandey, R.V. & Schlötterer C. 2011. PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*. **27**: 3435–3436.

Kofler, R. & Schlötterer, C. 2012 Gowinda: Unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics,* **28**: 2084–2085.

Kofler, R. & Schlötterer, C. 2014 A guide for the design of evolve and resequencing studies. *Molecular Biology and Evolution* **31**: 474–483.

Kolaczkowsi, B., Kern, A.D., Holloway, A.K. & Begun, D.J. 2011. Genomic differentiation between temperate and tropical australian populations of *Drosophila melanogaster*. *Genetics*. 187: 245–260.

Korneliussen, T.S., Albrechtsen, A. & Nielsen, R. 2013. ANGSD: Analysis of Next Generation Sequencing data. *BMC Bioinformatics*. **15**: 356.

Kronforst, M.R., Young, L.G., Kapan, D.D., McNeeley, C., O'Neill, R.J. & Gilbert, L.E. 2006. Linkage of butterfly mate preference and wing color preference cue at the genomic location of *wingless*. *PNAS*. **103**: 6575–6580.

Kosiol, C., Vinař, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R. & Siepel, A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genetics.* **4**: e1000144.

Koštál, V. 2011. Insect photoperiodic calendar and circadian clock: independence, cooperation, or unity? *Journal of Insect Physiology*. **57**: 538–556.

Kubatko & Degnan 2007

Kunte, K., Zhang, W., Tenger-Trolander, A., Palmer, D.H., Martin, A., Reed, R.D., Mullen, S.P. & Kronfrost, M.R. 2014. *doublesex* is a mimicry supergene. *Nature*. **507**: 229 – 232.

Küpper, C., Stocks, M., Risse, J.E., Remedios, N., Farrell, L.L., Mcrae, B., Morgan, T.C., Karlionova, N., Pinchuk, P., Verkuil, Y.I., Kitaysky, A.S., Wingfield, J.C., Piersma, T., Zeng, K., Slate, J., Blaxter, M., Lank, D.B. & Burke, T. 2016. A

supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics*. **48**: 79 – 83.

Lagisz, M., Wen, S-Y., Routtu, J., Klappert, K., Mazzi, D., Morales-Hojas, R., Schäfer, M.A., Vieira, J., Hoikkala, A., Ritchie, M.G. & Butlin, R.K. 2012. Two distinct genomic regions harbouring the *period* and *fruitless* genes affect male courtship song in *Drosophila montana*. *Heredity*. **108**: 602-608.

Laine V.N., Gossmann, T.I., Schachtschneider, K.M., Garroway, C.J., Madsen, O., Verhoeven, K.J.F., de Jager, V., Megens, H-J., Warren, W.C., Minx, P., Slate, J., Zeng, K., van Oers, K., Visser, M.E. & Groenen, M.A.M. 2016. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nature Communications*. **7**: 10474.

Lamichhaney, S., Martinez Barrio, A., Rafati, N., Sundström, G., Rubin, C-J., Gilbert, E.R., Berglund, J., Wetterbom, A., Laikre, L., Webster, M.T., Grabherr, M., Ryman, N. & Andersson, L. 2012. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *PNAS*. **109**: 19345-19350.

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

Landis, J.R., Heyman, E.R. & Koch, G.G. 1978. Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests. *International Statistical Review,* **46**: 237–254.

Lanfear, R., Ho, S.Y.W., Love, D. & Bromham, L. 2010. Mutation rate is linked to diversification in birds. *PNAS*. **107**: 20423-20428.

Langley, C.H., Stevens, K., Cardeno, C., Lee, Y.C.G., Schrider, D.R., Pool, John E., Langley, S.A., Suarez, C., Corbett-Detig, R.B., Kolaczkowski, B., Fang, S., Nista, P.M., Holloway, A.K., Kern, A.D., Dewey, C.N., Song, Y.S., Hahn, M.W. & Begun, D.J. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics*. **192**: 533–598.

Lankinen, P., Tyukmaeva, V.I. & Hoikkala, A. 2013. Northern *Drosophila montana* flies show variation both within and between cline populations in the critical day

length evoking reproductive diapause. *Journal of Insect Physiology*. **59**: 745–751.

Laturney, M. & Billeter, J.C. 2014. *Neurogenetics of female reproductive behaviors in drosophila melanogaster*. Elsevier.

Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D. 2012. Inference of population structure using dense haplotype data. *PLoS Genetics*. **8**: e1002453.

Lê, S., Josse, J. & Husson, F., 2008. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software.* **25**:1–18.

Lebreton, S., Grabe, V., Omondi, A.B., Ignell, R., Becher, P.G., Hansson, B.S., Sachse, S. & Witzgall, P. 2014. Love makes smell blind: mating suppresses pheromone attraction in *Drosophila* females via Or65a olfactory neurons. *Scientific Reports*. **4**: 7119.

Lee, J.E.A., Mitchell, N.C., Zaytseva, O., Chahal, A., Mendis, P., Cartier-Michaud, A., Parsons, L.M., Poortinga, G., Levens, D.L., Hannan, R.D. & Quinn, L.M. 2015. Defective Hfp-dependent transcriptional repression of *dMYC* is fundamental to tissue overgrowth in *Drosophila* XPB models. *Nature Communications*. **6**: 7404.

Lee, Y.W., Gould, B.A. & Stinchcombe, J.R. 2014. Identifying the genes underlying quantifying traits: a rationale for the QTN programme. *The Annals of Botany: Plants*. 6: plu004.

Lenski, R.E. 2011. Evolution in action: a 50,000-generation salute to Charles Darwin. *Microbe*. **6**: 30 – 33.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li, H., Hansdaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin., R. & 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. **25**: 2078–2079.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN]

Limousin, D., Streiff, R., Courtois, B., Dupuy, V., Alem, S. & Greenfield, M.D. 2012. Genetic architecture of sexual selection: QTL mapping of male song and female receiver traits in an acoustic moth. *PLoS ONE*. **7**: e44554.

Löytynoja, A. & Goldman, N. 2005. An algorithm for progressive multiple alignment of

sequences with insertions. *PNAS*. **102**: 10557–10562.

Lu, B., LaMora, A., Sun, Y., Welsh, M.J. & Ben-Shahar, Y. 2012. *ppk23*-dependent chemosensory functions contribute to courtship behavior in *Drosophila melanogaster*. **8**: e1002587.

Lund, S.P., Nettleton, D., McCarthy, D.J. & Smyth, G.K. 2012. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology,* **11**: Art8.

Lynch, M., Bost, D., Wilson, S., Maruki, T. & Harrison, S. 2014. Population-genetic inference from poole-sequencing data. *Genome Biology and Evolution.* **6:**1210–1218.

Machado, H.E., Bergland, A.O., O'Brien, K.R., Behrman, E.L., Schmidt, P.S. & Petrov, D.A. 2016. Comparative population genomics of latitudinal variation in *Drosophila simulans* and *Drosophila melanogaster*. *Molecular Ecology*. **25**: 723–740.

Mackay, T.F.C., Heinsohn, S.L., Lyman, R.F., Moehring, A.J., Morgan, T.J. & Rollman, S.M. 2005. The genetics and genomics of *Drosophila* mating behavior. *PNAS*. **102**: 6622-6629.

Mackay, T.F.C., Stone, E.A. & Ayroles, J.F. 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*. **10**: 565 – 577.

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

MacMillan, H.A, Ferguson, L. V., Nicolai, A., Donini, A., Staples, J. F. & Sinclair, B. J. 2015a. Parallel ionoregulatory adjustments underlie phenotypic plasticity and evolution of *Drosophila* cold tolerance. *Journal of Experimental Biology*. **218**: 423–432.

MacMillan, H.A., Andersen, J.L., Davies, S.A. & Overgaard, J. 2015b. The capacity to maintain ion and water homeostasis underlies interspecific variation in *Drosophila* cold tolerance. *Scientific Reports*. **5**: 18607

Magwene, P.M., Willis, J.H. & Kelley, J.K. 2011. The statistics of Bulk Segregant Analysis using next generation sequencing. *PLoS Computational Biology*. **7**: e1002255.

Mallet, J. 1989. The genetics of warning colour in peruvian hybrid zones of *Heliconius erato* and *H. melpomene*. *Proceedings of the Royal Society: B Biological Sciences*. **236**: 163 – 185.

Mank, J.E., Vicoso, B., Berlin, S. & Charlesworth, B. 2010. Effective population size and the faster-X effect: empirical results and their interpretation. *Evolution*. **64**: 663–674.

Mantel, N. & Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute,* **22**: 719–748.

Marcillac, F., Grosjean, Y. & Ferveur, J-F. 2005. A single mutation alters production and discrimination of *Drosophila* sex pheromones. *Proceedings of the Royal Society: Biological Science*. **272**: 303–309.

Marguerat, S. & Bähler, J. 2010. RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences*. **67**: 569 – 579.

Martin, O.Y., & Hosken, D.J. 2003. Costs and benefits of evolving under experimentally enforced polyandry or monogamy. *Evolution*. **57**: 2765–2772.

Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., Blaxter, M., Manica, A., Mallet, J. & Jiggins, C.D. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*. **23**: 1817-1828.

Martins, N.E., Faria, V.G., Nolte, V., Schlötterer, C. & Teixeira, L. 2014. Host adaptation to viruses relies on few genes with different cross-resistance properties. *Proceedings of the National Academy of Sciences,* **111**: 5938–5943.

Maston, G.A., Evans, S.K. & Green, M.R. 2006. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*. **7**: 29–59.

Matsui, H., Hunt, G.R., Oberhofer, K., Ogihara, N., McGowan, K.J., Mithraratne, K., Yamasaki, T., Gray, R.D. & Izawa, E-I. 2012. Adaptive bill morphology for enhanced tool manipulation in New Caledonian crows. Scientific Reports. **6**: 22776.

McBride, C.S. 2007. Rapid evolution of smell and tase receptor genes during host specialization in *Drosophila sechellia*. *PNAS*. **104**: 4996-5001

McCarthy, D.J., Chen, Y. & Smyth, G.K. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research,* **40**: 4288–4297.

McCue, M.E., Bannasch, D.L., Petersen, J.L., Gurr, J., Bailey, E., Binns, M.M., Distl, O., Guérin, G., Hasegawa, T., Hill, E.W., Leeb, T. Lindgren, G., Penedo, M.C.T., Røed, K.H., Ryder, O.A., Swinburne, J.E., Tozaki, T., Valberg, S.J., Vaudin, M., Lindblad-Toh, K., Wade, C.M. & Mickelson, J.R. 2012. A high density SNP array for the domestic horse and extant Perissodactyla: Utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genetics* **8**: e1002451.

McGraw, L.A., Clark, A.G. & Wolfner, M.F. 2008. Post-mating gene expression profiles of female *Drosophila melanogaster* in response to time and to four male accessory gland proteins. *Genetics*. **179**: 1395-1408.

McGraw, L.A., Gibson, G., Clark, A.G. & Wolfner, M.F. 2004. Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. *Current Biology*. **14**: 1509-1514.

McGregor, A.P., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D.G., Payre, F. & Stern, D.L. 2007. Morphological evolution through multiple *cis*-regulatory mutations at a single gene. *Nature*. **448**: 587-590.

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

McLeay, R.C. & Bailey, T.L. 2010. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. **11**: 165.

McLeay, R.C. & Bailey, T.L. 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**: 165.

Medina, F.S., Hunt, G.R., Gray, R.D., Wild, J.M. & Kubke, M.F. 2013. Perineural

satellite neuroglia in the telencephalon of New Caledonian crows and other Passeriformes: evidence of satellite glial cells in the cetral nervous system of healthy birds? *PeerJ*. **1**: e110.

Mendes, F.K. & Hahn, M.W. 2016. Gene tree discordance causes apparent substitution rate variation. *Systematic Biology*. **65**: 711-721.

Melcher, C. & Pankratz, M.J. 2005. Candidate gustatory interneurons modulating feeding behavior in the *Drosophila* brain. *PLoS Biology*. **3**: e305

Mero, I.-L., Lorentzen, A.R., Ban, M., Smestad, C., Celius, E.G., Aarseth, J.H., Myhr, K-M., Link, J., Hillert, J., Olsson, T., Kockum, I., Masterman, T., Oturai, A.B., Søndergaard, H.B., Sellebjerg, F., Saarela, J., Kemppinen, A., Elovaara, I., Spurkland, A., Dudbridge, F., Lie, B.A. & Harbo, H.F. 2010. A rare variant of the TYK2 gene is confirmed to be associated with multiple sclerosis. *European Journal of Human Genetics* **18**: 502–504.

Merrill, R.M., van Schooten, B., Scott, J.A. & Jiggins, C.D. 2011. Pervasive genetic associations between traits causing reproductive isolation in *Heliconius* butterflies. *Proceedings of the Royal Society: Biological Sciences*. **278**: 511–518.

Mirarab, S., Reaz, R., Bayzid, S., Zimmermann, T., Swenson, M.S. & Warnow, T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*. **30**: 541-548.

Mohammed, J., Flynt, A.S., Panzarino, A.M., Mondal, M.M.H., Siepel, A. & Lai, E.C. 2017. Deep experimental profiling of microRNA diversity, deployment, and evolution across the *Drosophila* genus. *BiorXiv*. doi: https://doi.org/10.1101/125997

Montgomery, S.L., Vorojeikina, D., Huang, W., Mackay, T.F.C., Anholt, R.R.H. & Rand, M.D. 2014. Genome-wide association analysis of tolerance to methylmercury toxicity in *Drosophila* implicates myogenic and neuromuscular developmental pathways. *PLoS ONE*. **9**: e110375

Morales-Hojas, R., Päällysaho, S., Vieira, C.P., Hoikkala, A. & Vieira, J. 2007. Comparative polytene chromosome maps of *D. montana* and *D. virilis*. *Chromosoma*. **116**: 21-27.

Moran, C. & Kyriacou, C. 2009. Functional neurogenomics of the courtship song of male *Drosophila melanogaster*. *Cortex.* **45**: 18–34.

Morgan, A., Ness, R.W., Keightley, P.D. & Colegrave, N. 2014. Spontaneous mutation accumulation in multiple strains of the green alga, *Chlamydomonas reinhardtii*. *Evolution*. **68**: 2589 – 2602.

Morley, K.I. & Montgomery, G.W. 2001. The genetics of cognitive process: candidate genes in Humans and animals. *Behaviour Genetics*. **31**: 511-531.

Nadeau, N.J., Ruiz, M., Salazar, P., Counterman, B., Medina, J.A., Ortiz-Zuazaga, H., Morrison A., McMillan, W.O., Jiggins, C.D. & Papa, R. 2014. Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome Research*. **24**: 1316 – 1333.

Nadeu, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., Quail, M.A., Joron, M., ffrench-Constant, R.H., Blaxter, M.L., Mallet, J. & Jiggins, C. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society: Biological Science*. **367**: 343-353.

Nakamura, M., Baldwin, D., Hannaford, S., Palka, J. & Montell, C. 2002. Defective proboscis extension response (DPR), a member of the Ig superfamily required for the gustatory response to salt. *The Journal of Neuroscience*. **22**: 3463–3472.

Nam, K., Mugal, C., Nabholz, B., Schielzeth, H., Wolf, J.B.W., Backström, N., Künstner, A., Balakrishnan, C.N., Heger, A., Ponting, C.P., Clayton, D.F. & Ellegren, H. 2010. Molecular evolution of genes in avian genomes. *Genome Biology*. **11**: R68.

Nasir, J., Floresco, S.B., Kusky, J.R.O., Diewert, V.M., Richman, J.M. Zeisler, J., Borowski, A., Marth, J.D., Phillips, A.G. & Hayden, M.R. 1995. Targeted disruption of Huntington's disease gene results in embryonic lethality and behaviour and morphological changes in heterozygotes. *Cell*. **81**: 811-823.

NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. **44**: D7–D10.

Neal, S.J., Karunanithi, S., Best, A., So, A.K-C., Tanguay, R.M., Atwood, H.L. & Westwood, J.T., 2006. Thermoprotection of synaptic transmisson in a *Drosophila* heat shock factor mutant is accompanied by increased expression of *Hsp83* and *DnaJ-1*. *Physiological Genomics*. **25**: 493–501.

Ness, R.W., Kraemer, S.A., Colegrave, N. & Keightley, P.D. 2015. Direct estimation of the spontaneous mutation rate uncovers the effect of drift and recombination in the *Chlamydomonas reinhardtii* plastid genome. *Molecula Biology and Evolution*. **33**: 800 – 808.

Neumüller, R.A., Richter, C., Fischer, A., Novatchkova, M., Neumüller, K.G. & Knoblich, J.A. 2011. Genome-wide analysis of self-renewal in *Drosophila* neural stem cells by transgenic RNAi. *Cell Stem Cell*. **8**: 580-593.

Neville, M.C., Nojima, T., Ashley, E., Parker, D.J., Walker, J., Southall, T., van de Sande, B., Marques, Ana C.Fischer, B., Brand, A.H.H., Russell, S., Ritchie, M.G., Aerts, S. & Goodwin, S.F. 2014. Male-specific fruitless isoforms target neurodevelopmental genes to specify a sexually dimorphic nervous system. *Current Biology*. **24**: 229-241.

Nicholson, G., Smith, A.V., Jónsson, F., Gústaffson, Ó., Stefánsson, K. & Donnelly, P. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B,* **64**: 695–715.

Noor, M.A.F. & Bennet, S.M. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*. **103**: 439-444.

Noor, M.F., Schug, M.D. & Aquadro, C.F. 2000. Microsatellite variation in populations of *Drosophila pseudoobscura* and *Drosophila perisimilis*. *Genetics Research*. **75**: 25-35.

Nyberg, K.G. & Machado, C.A. 2016. Comparative expression dynamics of intergentic long noncoding RNAs in the genus *Drosophila*. *Genome Biology and Evolution.* **8**: 1839–1858.

Oneal, E., Lowry, D.B., Wright, K.M., Zhu, Z. & Willis, J.H. 2014. Divergent population structure and climate associations of a chromosomal inversion polymorphism across the *Mimulus guttatus* species complex. *Molecular Ecology*. **23**: 2844-2860

Orozco-terWengel, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T. & Schlãtterer, C. 2012. Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular Ecology.* **21**: 4931-4941.

Orozco-terWengel, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T. & Schlötterer, C. 2012. Data from: Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Dryad Digital Repository*. doi: http://dx.doi.org/10.5061/dryad.60k68.2

Paolucci, S., Salis, L., Vermeulen, C.J., Beukeboom, L.W. & van de Zande, L. 2016. QTL analysis of the photoperiodic response and clinal distribution of *period* alleles in *Nasonia vitripennis*. **25**: 4805–4817.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20:** 289–290.

Pardo-Diaz, C., Salazar, C. & Jiggins, C.D. 2015. Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution.* **6**: 445-464.

Parker, D.J. & Vahed, K. 2009. The intensity of pre- and post-copulatory mate guarding in relation to spermatophore transfer in the cricket *Gryllus bimaculatus*. *Journal of Ethology*. **28**: 245–249.

Parker, D.J., Gardiner, A., Neville, M.C., Ritchie, M.G. & Goodwin, S.F. 2013. The evolution of novelty in conserved genes; evidence of positive selection in the *Drosophila* fruitless gene is localised to alternatively spliced exons. *Heredity*. **112**: 300-306.

Parker, D.J. Vesala, L., Ritchie, M.G., Laiho, A., Hoikkala, A. & Kankare, M. 2015. How consistent are the transcriptome changes associated with cold acclimation in two species *Drosophila virilis* group? *Herdity*. **115**: 13-21.

Parker, D.J., Cunningham, C.B., Walling, C.A., Stamper, C.E., Head, M.L., Roy-zokan, E.M., Mckinney, E.C., Ritchie, MG. & Moore, A.J. 2015. Transcriptomes of parents identify parenting strategies and sexual conflict in a subsocial beetle. *Nature Communications.* **6**: 8449.

Parker, D.J., Ritchie, M.G. & Kankare, M. 2016. Preparing for winter: The transcriptomic response associated with different day lengths in *Drosophila montana*. *G3* **6**: 1373–1381.

Parker, D.J., Vesala, L., Ritchie, M.G., Laiho, A., Hoikkala, A. & Kankare, M. 2015. How consistent are the transcriptome changes associated with cold acclimation in two species of the *Drosophila virilis* group. *Heredity*. **115**: 13–21.

Parker, D.J. et al., *in prep*. The *Drosophila montana* Genome

Parker, G. 1979. Sexual Selection and Sexual Conflict. Pp 123 – 166 in *Sexual Selection and Reproductive Competition in Insects* (M. Blum and A. Blum, Eds.). Academic Press, London.

Patel, M., Farzana, L., Robertson, L.K., Hutchinson, J., Grubbs, N., Shepherd, M.N. & Mahaffey, J.W. 2007. The appendage role of *disco* genes and possible implications on the evolution of the maggot larval form. *Developmental Biology.* **309**: 56–69.

Pegoraro, M., Zonato, V., Tyler, E.R., Fedele, G., Kyriacou, C.P. & Tauber, E. 2017. Geographical analysis of diapause inducibility in European *Drosophila melanogaster* populations. *Journal of Insect Physiology*. **98**: 238–244.

Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A. & Bejerano, G. 2013. Enhancers: five essential questions. *Nature Reviews Genetics*. **14**: 288–295.

Perry, J.C. & Rowe, L. 2014. The evolution of sexually antagonistic phenotypes. *Cold Spring Harbour Perspectives in Biology*. **7**: 1–18.

Pizzari, T. & Wedell, N. 2013. The polyandry revolution. *Philosophical Transactions of the Royal Society: Biological Sciences*. **368**: 20120041.

Poelstra, J.W., Vijay, N., Bossu, C.M., Lantz, H., Ryll, B., Müller, I., Baglione, V., Unneberg, P., Wikelska, M., Grabherr, M.G. & Wolf, J.B.W. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*. **344**: 1410–1414.

Polycansky, D. & Ellison, J. 1970. "Sex Ratio" in *Drosophila pseudoobscura*: spermiogenic failure. *Science*. **169**: 888-889.

Price, T.A.R., Bretman, A., Gradilla, A.C., Reger, J., Taylor, M.L., Campbell, A., Hurst, G.D.D. & Wedell, N. 2014. Does polyandry control population sex ratio via regulation of a selfish gene? *Proceedings of the Royal Society of London: Biology* **281**: 20133259.

Price, T.A.R., Hodgson, D.J., Lewis, Z., Hurst, G.D.D. & Wedell, N. 2008. Selfish genetic elements promote polyandry in a fly. *Science*. **211**: 1241-1243.

Price, T.A.R., Hurst, G.D.D. & Wedell, N. 2010. Polyandry prevents extinction. *Current Biology*. **20**: 471-475.

Price, T.A.R., Lewis, Z., Smith, D.T., Hurst, G.D.D. & Wedell, N. 2010. Sex ratio drive promotes sexual conflict and sexual coevolution in the fly *D. pseudoobscura*. *Evolution*. **64**: 1504-1509.

Price, T.A.R., Lewis, Z., Smith, D.T., Hurst, G.D.D. & Wedell, N. 2011. Remating in the laboratory reflects rates of polyandry in the wild. *Animal Behaviour*. **82**: 1381-1386.

Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. **155**: 477 – 959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. & Sham, P.C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**: 559–575.

Quinlan, A.R. & Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Development Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Available from: https://www.R-project.org/].

Radwan, J. & Babik, W. 2012. The genomics of adaptation. *Proceedings of the Royal Society: Biology*. **279**: 5024 – 5028.

Ramachandran, S., Rosenberg, N.A., Zhivotovsky, L.A. & Feldman, M.W. 2005. Robustness of the inference of human population structure: A comparison of X-chromosomal and autosomal microsatellites. *Human Genomics*. **1**: 87–97.

Rands, C.M., Darling, A., Fujita, M., Kong, L., Webster, M.T., Clabaut, C., Emes, R.D., Heger, A., Meader, S., Hawkins, M.B., Eisen, M.B., Teiling, C., Affourtit, J., Boese, B., Grant, P.R., Grant, B.R., Eisen, J.A., Abzhanov, A. & Ponting, C.P. 2013. Insights into the evolution of Darwin's finches from comparative analysis of the *Geospiza magnirostris* genome sequence. *BMC Genomics*. **14**: 95.

Rausher, M.D. & Delph, L.F. 2015. When does understanding phenotypic evolution require identification of the underlying genes? *Evolution*. **69**: 1655-1664.

Ravi Ram, K. & Wolfner, M.F. 2007. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. *Integrative and Comparative Biology*. **47**: 427-445.

Ravinet, M., Faria, R., Butlin, R.K., Galindo, J., Bierne, N., Rafajlovic, M., Noor, M.A.F., Mehlig, B. & Westram, A.M. in press. Interpreting the genomic landscape of

speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*. **XX**:XX-XX.

Rebeiz, M., Pool, J.E., Kassner, V.A., Aquadro, C.F. & Carroll, S.B. 2009. Stepwise modification of a modular enhancer underlies adaptation in a *Drosophila* population. *Nature*. **326**: 1663-1668.

Reed, L.K., Lee, K., Zhang, Z., Rashid, L., Hsieh, B., Deighton, N., Glassbrook, N., Bodmer, R. & Gibson, G. 2014. Systems genomics of metabolic phenotypes in wild-type *Drosophila melanogaster*. *Genetics*. **197**: 781-793.

Rendahl, K.G., Jones, K.R., Kulkami, S.J., Bagully, S.H. & Hall, J.C. 1992 The dissonance Mutation at the no-on-transient-A Locus of *D. melanogaster:* Genetic Control of Courtship Song and Visual Behaviors by a Protein with Putative RNA-binding Motifs. *The Journal of Neuroscience*. **72**: 390–407.

Richards, St., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., Couronne, O., Hua, S., Smith, M.A., Zhang, P., Liu, J., Bussemaker, H.J., van Batenburg, M., F., Howells, S.L., Scherer, S.E., Sodergren, E., Matthews, B.B., Crosby, M.A., Schroeder, A.J., Ortiz-Barrientos, D., Rives, C.M., Metzker, M.L., Muzny, D.M., Scott, G., Steffen, D., Wheeler, D.A., Worley, K.C., Havlak, P., Durbin, K.J., Egan, A., Gill, R., Hume, J., Morgan, M.B., Miner, G., Hamilton, C., Huang, Y., Waldron, L., Verduzco, D., Clerc-Blankenburg, K.P., Dubchak, I., Noor, M.A.F., Anderson, W., White, K.P., Clark, A.G., Schaeffer, S.W., Gelbart, W., Weinstock, G.M. & Gibbs, R.A. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Research*. **15**: 1-18.

Rinn, J.L. & Chang, H.Y. 2012. Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry*. **81**: 145–166.

Ritchie, M.G. & Butlin, R.K. 2014. The genetics of insect mating systems. Pp 59-77 in *The Evolution of Insect Mating Systems* (D.M. Shuker & Simmons, L.W. Eds). Oxford University Press, Oxford.

Ritchie, M.G., Saarikettu, M., Livingstone, S. & Hoikkala, A. 2013. Characterization of female preference functions for *Drosophila montana* courtship song and a test of the temperature coupling hypothesis. *Evolution*. **55**: 721–727.

Ritchie, M.G., Townhill, R.M. & Hoikkala, A. 1998. Female preference for fly song:

playback experiments confirm the targets of selection. *Animal Behaviour*. **56**: 713–717.

Rittschoff, C.C. & Robinson, G.E. 2014. Genomics: moving behavioural ecology beyond the phenotypic gambit. *Animal Behaviour*. **92**: 263-270.

Roch, S. & Steel, M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data can be statistically inconsistent. *Theoretical Population Biology*. **100**: 56-62.

Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. 2011. Identification of novel transcripts in annotated genomes using RNA-seq. *Bioinformatics*. **27**: 2325-2329.

Rockman, M.V. 2012. The QTN program and alleles that matter for evolution: All that's gold does not glitter. *Evolution*. **66**: 1-17.

Rodríguez, I. 2004. The *dachsous* gene, a member of the cadherin family, is required for Wg-dependent pattern formation in the *Drosophila* wing disc. *Development*. **131**: 3195–3206.

Rosato, E., Tauber, E. & Kyriacou, C.P. 2006. Molecular genetics of the fruit-fly circadian clock. *European Journal of Human Genetics*. **14**: 729-738.

Routtu, J., Mazzi, D., van der Linde, K., Mirol, P., Butlin, R. & Ritchie, M.G. 2007. The extent of variation in male song, wing and genital characters among allopatric *Drosophila montana* populations. *Journal of Evolutionary Biology*. **20**: 1591–1601.

Rutz, C. & St Clair, J.J.H. 2012. The evolutionary origins and ecological context of tool use in New Caledonian crows. *Behavioural Processes*. **89**: 153-165.

Rutz, C., Ryder, T.B. & Fleischer, R.C. 2012. Restricted gene flow and fine-scale population structuring in tool using New Caledonian crows. *Naturwissenschaften.* **99**: 313-320.

Rutz, C. Klump, B.C., Komarczyk, L., Leighton, R., Kramer, J., Wischnewski, S., Sugasawa, S., Morrissey, M.B., James, R., St. Clair, J.J.H., Switzer, R.A., Masuda, B.M, 2016. Discovery of species-wide tool use in the Hawaiian crow. *Nature*. **537**: 403-407.

Saarikettu, M., Liimatainen, J.O. & Hoikkala, A. 2005. The role of male courtship song in species recognition in *Drosophila montana*. *Behavior Genetics*. **35**: 257–263.

Sabb, F.W., Burggren, A.C., Higier, R.G., Fox, J., He. J., Parker, D.S., Poldrack, R.A., Chu, W., Cannon, T.D., Freimer, N.B. & Bilder, R.M. 2009. Challenges in

phenotype definition in the whole-genome era: multivariate models of memory and intelligence. *Neuroscience*. **164**: 88-107.

Sackton, T.B., Lazzaro, B.P., Schlenke, T.A., Evans, J.D., Hultmark, D. & Clark, A.G. 2007. Dynamic evolution of the innate immune system in *Drosophila*. **39**: 1461-1468.

Salazar-Jaramillo, L., Paspati, A., van de Zande, L., Vermeulen, C.J., Schwander, T. & Wertheim, B. 2014. Evolution of a cellular immune response in *Drosophila*: A phenotypic and genomic comparative analysis. *Genome Biology and Evolution*. **6**: 273-289.

Salih, D.A.M., Rashid, A.J., Colas, D., de la Torre-Ubieta, L., Zhu, R.P., Morgan, A.A., Santo, E.E., Ucar, D., Devarajan, K., Cole, C.J., Madison, D.V., Shamloo, M., Butte, A.J., Bonni, A., Josselyn, S.A. & Brunet, A. 2012. FoxO6 regulates memory consolidation and synaptic funciton. *Genes and Development.* **26**: 2780-2801.

Sambandan, D., Yamamoto, A., Fanara, J.J., Mackay, T.F.C. & Anhold, R.R.H. 2006. Dynamic genetic interactions determine odor-guided behavior in *Drosophila melanogaster*. *Genetics*. **174**: 1349–1363.

Santos, J., Pascual, M., Simões, P., Fragata, I., Rose, M.R. & Matos, M. 2013. Fast evolutionary genetic differentiation during experimental colonizations. *Journal of Genetics,* **92**: 183–194.

Saudou, F. & Humber, S. 2016. Review the biology of Huntingtin. *Neuron*. **89**: 910-926.

Schaeffer, S.W., Goetting-Minesky, M.P., Kovacevic, M., Peoples, J.R., Graybill, J.L., Miller, J.M., Kim, K., Nelson, J.G. & Anderson, W.W.2003. Evolutionary genomics of inversions in *Drosophia pseudoobscura*: evidence for epistasis. *PNAS*. **100**: 8319-8324.

Schäfer, M.A., Mazzi, D., Klappert, K., Kauranen, H., Vieira, J., Hoikkala, A., Ritchie, M.G. & Schlötterer, C. 2010. A microsatellite linkage map for *Drosophila montana* shows large variation in recombination rates, and a courtship song trait maps to an area of low recombination. *Journal of Evolutionary Biology*. **23**: 518–527.

Schäfer, M.A., Routtu, J., Vieira, J., Hoikkala, A., Ritchie, M.G. & Schlötterer, C. 2011. Multiple quantitative trait loci influence intra-specific variation in genital morphology between phylogenetically distinct lines of *Drosophila montana*. *Journal*

of Evoloutionary Biology. **24**: 1879–1886.

Schlichting, M., Menegazzi, P., Lelito, K. R., Yao, Z., Buhl, E., Dalla Benetta, E., Bahle, A., Denike, J., Hodge, J.J., Helfrich-Förster, C. & Shafer, O.T. 2016. A neural network underlying circadian entrainment and photoperiodic adjustment of sleep and activity in *Drosophila*. *Journal of Neuroscience*. **36**: 9084–9096.

Schlötterer, C., Kofler, R., Versace, E., Tobler, R. & Franssen, S.U. 2015. Combining experimental evolution with next-generation sequencing : a powerful tool to study adaptation from standing genetic variation. *Heredity,* **114**: 431–440.

Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*. **15**: 749–763.

Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**: 749–763.

Schmidt, P.S., Zhu, C-T., Das., J., Batavia, M., Yang, L. & Eanes, W.F. 2008. An amino acid polymorphism in the *couch potato* gene forms the basis for climatic adaptation in *Drosophila melanogaster*. **105**: 16207–16211.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jørgensen, J-E.Weigel, D. & Andersen, S.U. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*. **6**: 550-552.

Schneeberger, K. 2014. Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nature Reviews Genetics*. **15**: 662-676.

Schneider. A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G.H. & Graur, D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation and alignment. *Genome Biology and Evolution*. **1**: 114-118.

Schnorrer, F., Schönbauer, C., Langer, C.C.H., Dietzl, G., Novatchkova, M., Schernhuber, K., Fellner, M., Azaryan, A., Radolf, M., Stark, A., Keleman, K. & Dickson, B.J. 2010. Systematic genetic analysis of muscle morphogenenis and function in *Drosophila*. *Nature*. **464**: 287–291.

Schuster, S.C. 2008. Next-generation sequencing transforms today's biology. *Nature Methods*. **5**: 16-18.

Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L., Saetre, G-P., Bank, C., Brännström, A., Brelsford, A., Clarkson, C.S., Eroukhmanoff, F., Feder, J.L., Fischer, M.C., Foote, A.D., Franchini, P., Jiggins, C.D., Jones, F.C., Lindholm, A.K., Lucek, K., Maan, M.E., Marques, D.A., Martin, S.H., Matthews, B., Meier, J.I., Möst, M., Nachman, M.W., Nonaka, E., Rennison, D.J., Schwarzer, J., Watson, E.T., Westram, A.M. & Widmer, A. 2014. Genomics and the origin of species. *Nature Reviews Genetics*. **15**: 176 – 192.

Shaw, K.L. & Lesnick, S.C. 2009. Genomic linkage of male song and female acoustic preference QTL underlying a rapid species radiation. *PNAS*. **106**: 9737–9742.

Shenker, J.J., Sengupta, S.M., Joober, R., Malla, A., Chakravarty, M.M. & Lepage, M. 2017. Bipolar disorder risk gene *FOXO6* modulates negative symptoms in schizophrenia: a neuroimaging genetics study. *Journal of Psychiatry and Neuroscience*. **42**: 172-780.

Sheppard, P.M., Turner, J.R.G., Brown, K.S., Benson, W.W. & Singer, M.C.1985. Genetics and the evolution of Muellerian mimicry in *Heliconius* butterflies. *Philosophical Transactions of the Royal Society: Biology*. **308**: 433 – 610.

Shorter, J., Couch, C., Huang, W., Carbone, M.A., Pfeiffer, J., Anholt, R.R.H. & Mackay, T.F.C. 2015. Genetic architecture of natural variation in *Drosophila malanogaster* aggressive behaviour. *PNAS*. **112**: E3555–E3563.

Shuker, D.M. 2014. Sexual Selection Theory. Pp 20 – 41 in *The Evolution of Insect Mating Systems* (D.M. Shuker & L.W. Simmons Eds.). Oxford University Press, Oxford

Sinclair, B.J., Nelson, S., Nilson, T.L., Roberts, S.P. & Gibbs, A.G. 2007. The effect of selection for desiccation resistance on cold tolerance of *Drosophila melanogaster*. **32**: 322–327.

Singhal, S., Leffler, E.M., Sannareddy, K., Turner, I., Venn, O., Hooper, D.M., Strand, A.I., Li, Q., Raney, B., Balakrishnan, C.N., Griffith, S.C., Mcvean, G. & Przeworski, M. 2015. Stable recombination hotspots in birds. *Science*. **350**: 928-932.

Slatyer, R.A., Mautz, B.S., Backwell. P.R.Y. & Jennions, M.D. 2012. Estimating genetic benefits of polyandry from experimental studies: a meta-analysis. *Biological Reviews*. **87**: 1-33

Smeds L., Mugal, C.F., Qvarnström, A. & Ellegren, H. 2016. High-resolution mapping

of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genetics*. **12**: e1006044.

Smeds, L., Qvarnström, A. & Ellegren, H. 2016. Direct estimate of the rate *Genome Research*. **26**: 1211-1218.

Smith, C.L., Goldsmith, C-A.W. & Eppig, J.T. 2004. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*. **6**: R7.

Snook, R.R., Robertson, A., Crudgington, H.S. & Ritchie, M.G. 2005. Experimental manipulation of sexual selection and the evolution of courtship song in *Drosophila pseudoobscura*. *Behavior Genetics*. **35**: 245–255.

Snook, R.R., Brüstle L. & Slate, J. 2009. A test and review of the role of effective population size on experimental sexual selection patterns. *Evolution*. **63**: 1923-1933.

Snook, R.R. 2014. The evolution of polyandry. Pp 159–180 in *The Evolution of Insect Mating Systems* (D.M. Shuker & L.W. Simmons Eds.). Oxford University Press, Oxford

Sokal & Rohlf, 1969. *Biometry: The principles and practices of statistics in biological research.* W.H. Freeman and Company. San Francisco.

Sokal & Rohlf, 1981. *Biometry: The principles and practices of statistics in biological research.* W.H. Freeman and Company. New York

Sokolowski, M. 2001. *Drosophila*: genetics meets behaviour. *Nature Reviews: Genetics*. **2**: 879-890.

Stamatakis, A. 2014. RaXML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**: 1312-1313.

Stanziano, J.R., Sové, R.J., Rundle, H.D. & Sinclair, B.J. 2015. Rapid desiccation hardening changes the cuticular hydrocarbon profile of *Drosophila melanogaster*. *Comparative Biochemistry and Physiology, Part A*. **180**: 38–42

Stapely, J., Reger, J., Feulner, P.G.D., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A.D., Beckerman, A.P. & Slate, J. 2010. Adaptation genomics: the next generation. *Trends in Ecology and Evolution*. **25**: 705-712.

Steiner, C.C., Weber, J.N. & Hoekstra, H.E. 2007. Adaptive variation in beach mice produced by two interactinve pigmentation genes. *PLoS Biology*. **5**: e219.

Stern, D.L. & Frankel, N. 2013. The structure and evolution of *cis*-regulatory regions: the *shavenbaby* story. *Philosophila Transactions of the Royal Society: Biology*. **368**: 20130028.

Stern, D.L. & Orgogozo, V. 2008. The loci of evolution: how predictable is evolution? *Evolution*. **62**: 2155-2177.

Stinchcombe, J.R. & Hoekstra, H.E. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*. **100**: 158-170.

Stone, W.S., Guest, W.C. & Wilson, F.D. 1960. The evolutionary implications of the cytological polymorphism and phylogeny of the virilis group of *Drosophila*. *Genetics.* **46**: 350–361.

Storey, J. D. & Tibshirani, R. 2003. Statistical significance for genomewide studies. *PNAS*. **100**: 9440–9445.

Storey, J.D. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B Statistical Methodology* **64**: 479–498.

Storey, J.D., Bass, A., Dabney, A. & Robinson, D. 2015. qvalue: Q-value estimation for false discovery rate control. V 2.4.2

Stroumbakis, N.D., Li, Z. & Tolias, P.P. 1996. A homolog of human transcription factor NF-X1 encoded by the *Drosophila shuttle craft* gene is required in the embryonic central nervous system. *Molecular and Cellular Biology*. **16**: 192-201.

Sturtevant, A.H. & Dobzhansky, T.H. 1936. Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. *Genetics*. **22**: 448-450.

Swanson, W.J., Wong, A., Wolfner, M.F. & Aquadro, C.F. 2004. Evolutionary expressed seqence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics*. **168**: 1457–1465.

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

Takahashi, Y. 2015. Mechanisms and tests for geographic clines in genetic polymorphisms. *Population Ecology*. **57**: 355–362.

Tanaka, K.M., Hopfen, C., Herbert, M.R., Schlötterer, C., Stern, D., Masly, J.P., Mcgregor, A.P. & Nunes, M.D.S. 2015. Genetic architecture and functional

characterization of genes underlying the rapid diversification of male external genitalia between *Drosophila simulans* and *Drosophila mauritiana*. *Genetics*. **200**: 357-369.

Tapaltsyan, V., Charles, C., Hu, J., Mindell, D., Ahituv, N., Wilson, G.M., Black, B.L., Viriot, L. & Klein, O.D. 2016. Identification of novel *Fgf* enhancers and their role in dental evolution. *Evolution and Development*. **18**: 31-40

Tastan, Ö.Y., Maines, J.Z., Li, Y., Mckearin, D.M. & Buszczak, M. 2010. *Drosophila* Ataxin 2-binding protein 1 marks an intermediate step in the molecular differentiation of female germline cysts. *Development*. **137**: 3167–3176.

Tauber, E., Zordan, M., Sandrelli, F., Pegoraro, M., Osterwalder, N., Breda, C., Daga, A., Selmin, A., Monger, K., Benna, C., Rosata, E., Kyriacou, C.P. & Rodolfo, C. 2008. Natural selection favours a newly derived *timeless* allele in *Drosophila melanogaster*. Science. **316**: 1895–1899.

Taylor, M., Price, T.A.R. & Wedell, N. 2014. Polyandry in nature: a global analysis. *Trends in Ecology and Evolution*. **29**: 376-383.

Teramitsu, I. & White, S.A. 2006. *FoxP2* regulation during undirected singing in adult songbirds. *The Journal of Neuroscience*. **26**: 7390-7394.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*. **526**: 68-74.

The 1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*. **166**: 481-491.

The *Heliconius* Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*. **487**: 94 – 98.

Thimgan, M.S., Seugnet, L., Turk, J. & Shaw, P.J. 2015. Identification of genes associated with resilience/vulnerability to sleep deprivation and starvation in *Drosophila*. *Sleep*. **38**: 801–814.

Thistle, R., Cameron, P., Ghorayshi, A., Dennison, L. & Scott, K. 2012. Contact chemoreceptors mediate male-male repulsion and male-female attraction during *Drosophila* courtship. *Cell*. **149**: 1140–1151.

Throckmorton, L. 1982. The virilis species group. Pp. 227–296 in *The biology and genetics of Drosophila* (M. Ashburner, H. Carson & J. Thompson Eds.). Jr. Academic Press, London.

Thumm, M. & Kadowaki, T. 2001. The loss of *Drosophila APG/AUT2* function modifies the phenotypcs of *cut* and Notch signalling pathway mutants. *Molecular Genetics and Genomics*. **266**: 657–663.

Tinbergen, N. 1963. On aims and methods of ethology. *Zeitschrift für Tierpsychologie*. **20**: 410 – 433.

Toda, H., Zhao, X. & Dickson, B.J. 2012. The *Drosophila* female aphrodisiac pheromone activates *ppk23*$^{+}$ sensory neurons to elicit male courtship behaviour. *Cell Reports*. **1**: 599–607.

Topa, H., Jónás, Á., Kofler, R., Kosiol, C. & Honkela, A. 2015. Gaussian process test for high-throughput sequencing time series: applications to experimental evolution. *Bioinformatics*. **31**: 1762–1770.

Trapnell, C. Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L. 2010. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. **28**: 511-515.

Travisano, M. & Shaw, R.G. 2013. Lost in the map. *Evolution*. **67**: 305–314.

Tregenza, T., Wedell, N. & Chapman, T. 2006. Sexual conflict: a new paradigm? *Philosophical Transactions of the Royal Society: Biological Sciences*. **361**: 229–234.

Trivers, R., 1972. Parental investment and sexual selection. Pp 52–95 in *Sexual Selection and the Descent of Man* (B.G. Campbell, Ed.). Aldine, Chicago.

Troscianko, J., von Bayern, A.M.P., Chappell, J., Rutz, C. & Martin, G.R. 2012. Extreme binocular vision and a straight bill facilitate tool use in New Caledonian crows. *Nature Communications*. **3**: 1110.

Tsai, Y-C., Chiang, W., Liou, W., Lee, W-H., Chang, Y-W., Wang, P-Y., Li, Yi-C., Tanaka, T., Nakamura, A. & Pai, L-M. 2014. Endophilin B is required for the *Drosophila* oocyte to endocytose yolk downstream of Oskar. *Development*. **141**: 563–573.

Tsuda, M. & Aigaki, T. 2016. Evolution of Sex-Peptide in *Drosophila*. *Fly*. **10**: 172-177.

Tsuda, M., Peyre, J-B., Asano, T. & Aigaki, T. 2015. Visualizing molecular functions and cross-species activity of sex-peptide in *Drosophila*. *Genetics*. **200**: 1161-1169.

Tykmaeva, V.I., Veltsos, P., Slate, P., Gregson, E., Kauranen, H., Kankare, M., Ritchie,

M.G., Butlin, R.K. & Hoikkala, A. 2015. Localization of quantitative trait loci for diapause and other photoperiodically regulated life history traits important in adaptation to seasonally varying environments. *Molecular Ecology.* **24**: 2809–2819.

Tyukmaeva, V.I., Salminen, T.S., Kankare, M., Knott, K.E. & Hoikkala, A. 2011. Adaptation to a seasonally varying environment: a strong latitudinal cline in reproductive diapause combined with high gene flow in *Drosophila montana*. *Ecology and Evolution*. **1**: 160–168.

Ugrankar, R., Berglund, E., Akdemir, F., Tran, C., Kim, M.S., Noh, J., Schneider, R., Ebert, B. & Graff, J,M. 2015. *Drosophila* glucome screening identifies Ck1alpha as regulator of mammalian glucose metabolism. *Nature Communications*. **6**: 7102

Usha, N. & Shashidhara, L.S. 2010. Interaction between Ataxin 2-binding protein 1 and Cubitus-interruptus during wing development in *Drosophila*. *Developmental Biology*. **341**: 389–399.

van Dongen 2000. Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht.

Veltsos, P., Fang, X., Cossins, A.R. Snook, R.R. & Ritchie, M.G. Mating system manipulation and the evolution of sex-biased gene expression in Drosophila. *Nature Communications* (In revision).

Vesala, L. & Hoikkala, A. 2011. Effects of photoperiodically induced reproductive diapause and cold hardening on the cold tolerance of *Drosophila montana*. *Journal of Insect Physiology*. **57**: 46–51.

Vesala, L., Salminen, T.S., Kankare, M. & Hoikkala, A. 2012a. Photoperiodic regulation of cold tolerance and expression of regucalcin gene in Drosophila montana. Journal of Insect Physiology. 58: 704–709.

Vesala, L., Salminen, T.S., Koštál, V., Zahradníčková, H. & Hoikkala, A. 2012b. *Myo*-inositol as a main metabolite in overwintering flies: seasonal metabolomic profiles and cold stress tolerance in a northern drosophilid fly. *The Journal of Experimental Biology*. **215**: 2891–2897.

Vesala, L., Salminen, T.S., Laiho, A., Hoikkala., A. & Kankare, M. 2012c. Cold tolerance and cold-induced modulation of gene expression in two *Drosophila virilis* group species with different distributions. *Insect Molecular Biology*. **21**: 107–118.

Vicoso, B. & Charlesworth, B. 2006. Evolution on the X chromosome: unusual patterns

and processes. *Nature Reviews Genetics*. **7**: 645–653.

Vicoso, B. & Charlesworth, B. 2009. Effective population size and the faster-X: an extended model. *Evolution*. **63**: 2419-2426.

Vieira, F.G., & Rozas, J. 2011. Comparative genomics of the odorant-binding protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biology and Evolution.* **3**: 476–490

Vieira, F.G., Sánchez-Gracia, A. & Rozas, J. 2007. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. *Genome Biology*. **8**: R235

Vigoder, F.M., Parker, D.J., Cook, N., Tournière, O., Sneddon, T. & Ritchie, M.G. 2016. Inducing cold-sensitivity in the frigophilic fly *Drosophila montana* by RNAi. *PLoS ONE.* **11**: e0165724.

Vijay, N., Bossu, C.M., Poelstra, J.W., Weissensteiner, M.H., Suh, A., Kryukov, A. & Wolf, J.B.W. 2016. Evolution of heterogenous genome of differentiation across multiple contact zones in a crow species complex. *Nature Communications*. **7**: 13195.

Vines, T.H., Dalziel, A.C., Albert, A.Y.K., Veen, T., Schulte, P.M. & Schluter, D. 2016. Cline coupling and uncoupling in a stickleback hybrid zone. *Evolution*. **70**: 1023–1038.

Wallace, A.G., Detweiler, D. & Schaeffer, S.W. 2011. Evolutionary history of the third chromosome gene arrangements of *Drosophila pseudoobscura* inferred from inversion breakpoints. *Molecular Biology and Evolution*. **28**: 2219–2229.

Wang, J. 2005. Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society: Biological Sciences*. **360**: 1395-1409

Wang, N., Leung, H.-T., Pak, W.L., Carl, Y.T., Wadzinski, B.E. & Shieh, B.-H. 2008. Role of protein phosphatase 2A in regulating the visual signaling in *Drosophila*. *The Journal of Neuroscience*. **28**: 1444–1451.

Wedell, N. 2013. The dynamic relationship between polyandry and selfish genetic elements. *Philosophical Transactions of the Royal Society: Biological Sciences*. **368**: 20120049.

WellenreutherM., Rosenquist, H., Jaksons, P. & Larson W, K. 2017. Local adaptation along an environmental cline in a species with an inversions polymorphism. *Journal of Evolutionary Biology*. **30**: 1068-1077.

Weng, M-P. & Liao, B.Y. 2010. MamPhEA: a web tool for mammalian phenotype enrichment analysis. *Bioinformatics*. **26**: 2212-2213.

Weng, M-P. & Liao, B.Y. 2011. DroPhEA: Drosophila phenotype enrichment analysis for insect functional genomics. *Bioinformatics* **27**: 3218–3219.

Weng, R., Chin, J.S.R., Yew, J.Y., Bushati, N. & Cohen, S.M. 2013. *miR-124* controls male reproductive success in *Drosophila*. *eLife* **2**: e00640

Werner, T., Koshikawa, S., Williams, T.M. & Carroll S.B. 2010. Generation of a novel wing colour pattern by the Wingless morphogen. *Nature*. **464**: 1143–1148.

Wiley, C., Ellison, C.K. & Shaw, K.L. 2011. Widespread genetic linkage of mating signals and preferences in the Hawaiian cricket *Laupala*. *Proceedings of the Royal Society: Biological Sciences*. **279**: 1203–1209.

Wilkinson, G.S., Breden, F., Mank, J.E., Ritchie, M.G., Higginson, A.D., Radwan, J., Jaquiery, J., Salzburger, W., Arriero, E., Barribeau, S.M., Phillips, P.C., Renn, S.C.P. & Rowe, L. 2015. The locus of sexual selection: moving sexual selection studies into the post-genomics era. *Journal of Evolutionary Biology*. **28**: 739-755.

Williams, J.A. & Sehgal A. 2001. Molecular components of the circadian system in *Drosophila*. *Annual Review of Physiology*. **63**: 729–755.

Williams, T.M. & Carroll, S.B. 2009. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nature Reviews Genetics* **10**: 797–804.

Wittkopp, P.J. & Kalay, G. 2012. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*. **13**: 59-69.

Wolf, J.B.W. & Ellegren, H. 2016. Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*. **18**: 87–100.

Wolf, J.B.W. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular Ecology Resources*. **13**: 559 – 572.

Woolf, B. 1955. On Estimating the Relation Between Blood Group and Disease. *Annals of Human Genetics,* 19: 251–253.

Wright A.E., Harrison, P.W., Zimmer, F., Montgomery, S.H., Pointer, M.A. & Mank,

J.E. 2015. Variation in promiscuity and sexual selection drives avian rate of faster-Z evolution. *Molecular Ecology*. **24**: 1218–1235.

Wu., P, Jiang., T-X., Shen, J-Y., Widelitz, R.B. & Chuong, C-M. 2006. Morphoregulation of avian beaks: comparative mapping of growth zone activities and morphological evolution. *Development Dynamics*. **235**: 1400-1412.

Xu, T., Gu, L., Schachtschneider, K.M., Liu, X., Huang, W., Xie, M. & Hou, S. 2012. *PLoS ONE*. **9**: e107574.

Yang, J., Jiang, H., Yeh, C.T., Yu, J., Jeddeloh, J.A., Nettleton, D. & Schnable, P.S. 2015. Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. *The Pland Journal*. **84**: 587-596.

Yang, Z.1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*. **15**: 568 – 573.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*. **24**: 1586 – 1591.

Yapici, N., Kim, Y-J., Ribeiro, C. & Dickson, B.J. 2008. A receptor that mediates the post-mating switch in *Drosophila* reproductive behaviour. *Nature*. **451**: 33-38.

Zayed, A. & Robinson, G.E. 2012. Understanding the relationship between brain gene expression and social behavior: lessons from the honey bee. *Annual Review of Genetics*. **46**: 591-615.

Zbinden, M., Haag, C.R. & Ebert, D. 2008. Experimental evolution of field populations of *Daphnia magna* in response to parasite treatment. *Journal of Evolutionary Biology*. **21**: 1068–1078.

ZeBRA: A Zebra Finch Expression Atlas, [Accessed 06.06.2017] url: http://zebrafinchatlas.org.

Zelle, K., Lu, B., Pyfrom, S.C. & Ben-Shahar, Y. 2013. The genetic architecture of degenerin/epithelial sodium ion channels in *Drosophila*. *G3*. **3**: 441–450.

Zhang, Y., Nielsen, R. & Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*. **22**: 2472 – 2479.

Zhong, W., Mcclure, C.D., Evans, C.R., Mlynski, D.T., Immonen, E., Ritchie, M.G. & Priest, N.K. 2013. Immune anticipation of mating in *Drosophila*: *Turandot M*

promotes immunity against sexually transmitted fungal infections. *Proceedings of the Royal Society: Biological Sciences*. **280**: 20132018.

Zhu, L.J., Christensen, R.G., Kazemian, M., Hull, C.J., Enuameh, M.S., Basciotta, M.D., Brasefield, J.A., Zhu, C., Asriyan, Y., Lapointe, D.S., Sinha, S., Wolfe, S.A. & Brodsky, M.H. 2011. FlyFactorSurvey: A database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Research* **39**: 111–117.

Zuk, M. & Belanger, S.L. 2014. Behavioral ecology and genomics: new directions, or just a more detailed map? *Behavioral Ecology*. **25**: 1277 – 1282.

Appendix A: Standard phenol-chloroform protocol for DNA extraction.

This protocol is developed for pooled genomic DNA from 40 flies per sample. Samples stored in ethanol from collection.

Materials:
Insect Extraction Buffer (IEB) heated until clear (37-55^{o}C)
Proteinase-K
epicentre "Riboshredder"
Phenol:chloroform mix
Chloroform
100% ethanol (EtOH) (ice cold -20^{o}C)
PCR water (DNAse free)

Cell Lysis Steps:
1. Dry samples on paper towel.
2. Place samples (40 flies) in a 1.5ml eppendorf tube with 400 ul of IEB and homogenize with an electric homogenizer.

3. Transfer solution to a falcon tube, add 1200 ul IEB and 20 ul proteinase-K. Mix sample well by vortexing. Incubate over night at 55^{o}C.

RNA Removal
4. Transfer 1000 ul of solution from 3. into new 1.5ml eppendorf tubes (@ 500 ul per eppendorf tube)
5. Add 0.5 ul of Riboshredder and incubate at 37^{o}C for 30 minutes.

The above steps give 2 samples from each pool of 40 individuals. The DNA cleaning steps are performed on all samples.

DNA Cleaning
6. Add 500ul[1] phenol-chloroform[2] to the sample (in fume cupboard, wearing two pairs of gloves)
7. Ensure that eppendorf tubes are tightly closed and mix contents by shaking until an emulsion forms.

8. Centrifuge at 10,000 rpm for 10 minutes at room temperature. Check that organic and aqueous phases are well separated.

9. Remove the aqueous supernatant (350 ul) to a fresh 1.5ml eppendorf tube. Be careful not to take any of the phenol layer.

Several phenol:chloroform washes (up to 3) can be performed if there is any doubt as to the cleanliness.

10. Add 350ul[1] chloroform[2] and repeat steps 9-11. At step 9: remove 300 ul of aqueous supernatant
Chloroform binds any residual phenol to remove it from the sample. Again, several chloroform washes can be performed to remove the residual phenol.

11. Add 1ml of ice cold 100% EtOH. Mix sample by inverting and leave to precipitate for 2 hrs at -20^{o}C (sample should be clear to slightly cloudy at this stage)

12. Centrifuge at 13,000 rpm for 5 mins to condense DNA in pellet. Ideally at 4^{o}C
13.Pipette off the EtOH[3] without disrupting the pellet. First with P1000 pipette, then with P20 pipette.
14. Add 500ul EtOH[3] (70%). Spin at 13,000 rpm at 4^{o}C. Repeat step 13. Evaporate EtOH by spinning in vacuum "extractor" for 5 minutes.
14. Add 200ul of PCR water or TE buffer and incubate at 37^{o}C for 1hr
15. Remove 20ul of samples for NanoDrop and running on a gel. Store the main sample in the fridge.

[1] ratio of sample:phenol-chloroform (step 6) and sample:chloroform (step 10) should always be 1:1
[2] Stored in fridge. Phenol-chloroform is covered and protected from light.
[3] Doesn't need to be ice cold.