INVESTIGATING THE MOLECULAR BASIS OF ADAPTATION AND SPECIATION IN DIVERGENT POPULATIONS

Gilbert Smith

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

2013

INVESTIGATING THE MOLECULAR BASIS OF ADAPTATION AND SPECIATION IN DIVERGENT POPULATIONS

GILBERT SMITH

THIS THESIS IS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

21ST SEPTEMBER 2012

DECLARATIONS

1. Candidate's declaration:

I, Gilbert Smith, hereby certify that this thesis, which is approximately 55,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in October 2008 and as a candidate for the degree of Ph.D in October 2009; the higher study for which this is a record was carried out in the University of St Andrews between 2008 and 2012.

Date: 05/04/2013 signature of candidate:

2. Supervisor's declaration:

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

Date: 5/4/13 signature of supervisor: _____

3. Permission for electronic 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 electronic publication of this thesis:

Access to printed copy and electronic publication of thesis through the University of St Andrews.

Date: 05/04/2013 signature of candidate: signature of supervisor: _____

ACKNOWLEDGEMENTS

There are many people that I wish to thank but first I must begin by thanking my supervisor, Prof Mike Ritchie. Mike took me on as a research assistant and later gave me the opportunity to become his student. I have learnt innumerable valuable lessons from Mike over the years and have grown as an academic due to his support and generosity, particularly in allowing me to follow my own ideas, sending me to conferences and introducing me to many of his academic collaborators.

I would like to thank the whole of the Ritchie lab, old and new, for all their support and interesting discussions. This thanks extends to the entire of the Center for Evolution, Genes and Genomics. The lab chats were always illuminating and particularly useful for practicing conference talks. My thanks also to Daniel Barker who provided much useful bioinformatic advice over the years, and comments on my thesis.

Thanks must also go to several collaborators. William Etges, from the University of Arkansas, took me under his wing during fieldwork in Mexico and taught me a great deal about the ecology of *Drosophila mojavensis*. Bill has been amazing in handling the flies for my entire project, and has consistently given thorough and very useful comments on my manuscripts and project ideas. Thanks to Konrad Lohse from the University of Edinburgh for a lot of useful advice on analyses and manuscripts, and for being a supportive friend. Thank you to Tino Macías Garcia for supervision and support, and Yoli Saldívar for advice and support on all things fishy.

I would also like to thank Lianne Baker who has been a wonderful secretary of Dyers Brae House. Lianne is always willing to help and was kind enough to handle my submission-from-afar.

My deepest thanks go to Tanya Sneddon, without whom none of the wet-lab work would have been possible. Tanya is an amazing technician with a wide-range of wet-lab skills and I cannot count the number of times she has helped me out of a PCR black hole.

Thanks also go to the Center for Genomic Research at the University of Liverpool for assistance in all aspects of high throughput sequencing.

My last and most profound thanks go to my family. First, to Laurel Fogarty who has supported me throughout my PhD and especially during the writing of my thesis. Thank you Laurel for always being there and knowing what to say. My deepest thanks to Tom Smith, Frank Norman and Mike and Madeline Eve, who were there when I needed them. Lastly, and importantly, I must thank my parents and sisters Alan, Anthea, Cathy, Rosy and Bella. Thank you all for unerring support over many years, without you I would not be where I am today.

CONTENTS

ABSTRACT

The creation of biodiversity involves the evolution of new species. Recent trends in the study of speciation have increased the emphasis on the role of ecology in adaptation and the evolution of reproductive isolation. This includes examining the relative contributions of different types of selection, the role of gene flow and the genomic changes that occur during ecological speciation. The search for speciation genes continues, however our growing knowledge of how the genome translates into phenotypes means we should now consider a broader molecular basis of speciation, which includes genetic, transcriptomic and potentially epigenetic variation that contribute to phenotypic variation. This thesis addresses the molecular basis of speciation by using three different complementary methods to examine the early stages of ecological speciation and the evolution of premating reproductive isolation between two incipient species of the cactophilic fly, *Drosophila mojavensis*. First, the genetic basis was examined through the sequencing of two candidate genes underlying reproductive isolation (Chapter 2). Second, the historical biogeography of population divergence was uncovered using multiple sequenced loci (Chapter 3). Lastly, gene expression across the whole transcriptome associated with phenotypic plasticity and mating success was assessed (Chapter 4). Further, the role of epigenetic imprinting in the population divergence of a freshwater fish, *Girardinichthys multiradiatus*, was examined through sequencing of a well known gene involved in sexual conflict (Chapter 5). These studies find that uncovering the genetic variation underlying speciation is difficult, especially when there is extensive phenotypic plasticity. Further, gene expression plasticity may play an important role in the evolution of premating isolation, and this includes a role for epigenetic mechanisms of gene expression. Additionally, it is important to assess the demographic scenario of population divergence to put into context the ecological and functional data on divergent groups. Through these studies this thesis examines the genetic, expression and epigenetic variation associated with on-going population divergence, and emphasises the need to consider the potential role of the full range of gene expression changes during ecological speciation.

CHAPTER 1:

GENERAL INTRODUCTION

Speciation proceeds with the evolution of reproductive barriers between divergent populations (Noor & Feder 2006; Wu & Ting 2004). Although many advances in speciation research have been made in the last one hundred and fifty years, many of the key questions asked by Charles Darwin about the origin of species, "that mystery of mysteries" (Darwin 1859), still remain unanswered (Butlin *et al*. 2012; Mank 2009). Diverging populations accumulate genetic changes that, over time, might lead to genetic incompatibilities and changes in reproductive characters, resulting in a reduction in gene flow and reproductive isolation. Exactly how these isolating factors accumulate or how reproductive isolating mechanisms evolve remains unclear, and this is currently one of the key unanswered questions in speciation research today. Another key question that has received much attention is whether species arise as a consequence of natural selection or through genetic drift (and chance events e.g. mutation, chromosomal inversions; Dobzhansky 1937; Muller 1939). However, there is now growing evidence to suggest that many species have evolved through some form of selective process (Schluter 2009 and examples therein). It is worth noting, however, that genetic drift is likely to have played a role in speciation to some degree, and it is more pertinent to assess the relative roles of drift and selection in population divergence.

The definition or delimitation of a species has been a controversial topic in speciation research for many years, and remains so today. Upwards of 22 species concepts have been defined over the years, with most being developed in the last few decades (Mayden 1997). Each new definition has been coined, it would seem, to address problems or circularities in previous definitions (Mallet 1995). However, no single definition encompasses every aspect of a biological species and thus each tends to focus on slightly different biological properties (De Queiroz 2005). For example

the genetic species concept is useful when comparing genetic distinctiveness to identify putative species, and is defined as a group of genetically compatible interbreeding natural populations that is genetically isolated from other such groups (Dobzhansky 1950; Simpson 1943; definition from Baker & Bradley 2006). This definition is useful for identifying genetic groups, however genetic distinctiveness in this case is used to *infer* reproductive isolation rather than directly assessing their cause and effect relationship (Ferguson 2002). Similarly, the BSC emphasises reproductive isolation, the ecological species concept focuses on ecological niches and so on. The differing species concepts and criteria for species delimitation have been reviewed and revised many times (e.g. De Queiroz 2005, 2007, 2011; Mallet 2001; Wiens 2007) and it seems that a researcher must chose and adhere to one particular definition that is applicable to the question they address. Reproductive isolation is the key identification for a 'good species' under the BSC, which is often seen as the gold standard to which putative (sexually reproducing) species must be compared. The BSC is defined as groups of actually or potentially interbreeding natural populations that are reproductively isolated from other such groups. However, the BSC has been criticised primarily for not including well differentiated and generally accepted 'good species' than can in fact interbreed (Hausdorf 2011).

To understand the process of speciation we must examine groups of individuals along a continuum of divergence and the existence of organisms at various stages of this continuum is a principle reason for the lack of a single coherent species concept (De Queiroz 2005; Hausdorf 2011; Mallet 2001). The question then becomes *when* do we consider diverging populations to be species? As such I take the BSC as the broad standard for defining species, acknowledging that reproductive isolation can be imperfect even between 'good species' (Coyne & Orr 1998; Harrison

1998; Rundle & Schluter 2004; Schluter 2001), yet I do not adhere to any one definition or attempt to explain any observed divergence under any specific definition. I take a hierarchical approach when describing groups of individuals in which sub-populations, populations and species may be a progression of differentiation with a corresponding decrease in levels of gene flow and quantitative increase of reproductive isolation in nature (Schluter et al, 2001; Coyne and Orr, 1998; Harrison, 1998).

Reproductive isolation

Reproductive isolation has been studied for a number of years, yet there are still many unanswered questions as to how it evolves. For example, what kinds of traits cause reproductive isolation and what forms of selection act on them (Boughman *et al*. 2005)? How often is reproductive isolation a by-product of adaptation versus direct selection for isolation (Rundle & Schluter 2004)? Sexual selection on reproductive traits may help drive divergence; if so, can it act alone or only in conjunction with other sources of disruptive selection as some theoretical models seem to suggest (Kirkpatrick & Ravigne 2002; Lande 1981; M'Gonigle *et al*. 2012; Ritchie 2007). Further, we know that isolation has a genetic basis but at what level of genome-wide reproductive isolation do we define populations as species (Wu 2001)?

Reproductive isolation occurs through genetic changes within populations that create isolating barriers and restrict gene flow (Coyne & Orr 1989). It is difficult to define exactly when two groups can be considered reproductively isolated. If hybrids can arise between two populations, gene flow may still occur meaning the groups are effectively still one gene pool (Wu & Ting 2004). Thus, it is difficult to say exactly how much of a genome must be reproductively isolated before we define divergent

groups as species. This continuum of genetic differences can underlie different modes of reproductive barrier and, at the most extreme (i.e. a 'good' species), the whole genome of a species could be considered a single coadapted unit, incompatible with any other species (Wu 2001). Reproductive barriers to gene flow can be classified as premating; postmating, prezygotic; or postzygotic (Coyne & Orr 2004). Premating barriers include features that might prevent mating between species through behavioural, ecological or mechanical methods. Examples include mate recognition and preference systems, temporal differences in reproduction and incompatibilities in reproductive structures. Postmating, prezygotic isolating barriers can include copulatory isolation or gametic isolation, in which sperm is restricted when entering the egg due to incompatible surface proteins. Postzygotic isolating barriers include fitness reduction of hybrids or hybrid inviability and sterility. Often multiple modes of reproductive isolation can exist between species, leading to problems in knowing which one caused the initial speciation event.

Coyne and Orr (1989) found that prezygotic isolation in sympatric populations of *Drosophila* species evolved at a faster rate than postzygotic isolation, however in allopatry these rates were about equal. This was thought to be due to reinforcement where, in sympatric populations, the presence of hybrids creates a selective pressure for assortative mating through natural selection against hybrids (Servedio 2001). Sexual isolation between populations, defined as premating isolation caused by courtship traits and associated preferences (Ritchie 2007), is thought to represent a first step in the gradual evolution of reproductive isolation (Mackay *et al*. 2005) and is also thought to be a common cause of reproductive isolation in animal species. For example, in birds speciation often begins with behavioural premating isolating barriers, with postmating barriers arising later (Grant & Grant 1997). Thus it is

possible that premating isolation evolves early and fairly rapidly between diverging populations.

Although recognised as an important barrier in maintaining genetic and phenotypic differences between species, the evolutionary origins of sexual isolation remain unclear (Boake *et al*. 1997; Ting *et al*. 2001). Reinforcement is often cited as a cause, yet when presented with a case of sexual isolation it is very difficult to determine if the cause was reinforcement or ecological adaptation and indirect selection for reproductive traits (Servedio & Noor 2003). Even when considering reinforcement, direct selection on female preferences might be the primary driving force of isolation, rather than the more traditional definition of selection against hybrids (Servedio 2004). Behavioural isolation (or species recognition) is often considered as part of a continuum of mate choice, and studies have shown that isolation can evolve through sexual selection for mate recognition characters within groups (Lande & Kirkpatric 1988; Turner & Burrows 1995). There are several theories of sexual selection (including the "Good-genes" hypothesis, the "sexy-son" hypothesis and antagonistic sexual selection; Weatherhead & Robertson 1979; Zahavi 1975) and these models are thought to be a continuum of a single process (Kokko *et al*. 2002). However, Paterson (1985; 1993) recognized that sexual selection and species recognition might be different systems, coining the term "specific-mate recognition system". For example, a novel trait in *Drosophila heteroneura* was shown to be involved in sexual selection but not in the discrimination of a closely related species, *Drosophila silvestris* (Carson *et al*. 1989). Although the relationship between intraspecific mate recognition systems and species recognition systems is still under some debate it is possible the same processes create both; with continued divergence comes species recognition. However, separating those traits involved in species

recognition from those purely under sexual selection may be difficult as, in the case of many *Drosophila* species, courtship involves several different cues. Here I focus on premating, behavioural isolation as an important early stage of the evolution of reproductive isolation in incipient (newly emerging) species, occasionally touching on postmating barriers when relevant.

The changing face of speciation research

Evidence shows that speciation might occur under many different conditions depending on the mode of speciation, the balance of interacting evolutionary forces, the strength of these forces, and the initial conditions (Gavrilets 2003). For example disruptive selection might come from environmental selection and sexual selection acting in unison (Noor 1997; Ritchie 2007). Alternatively, sexual selection is considered a potent process alone (Ritchie 2007; West-Eberhard 1983), and recent work shows that strong sexual selection might be a primary driver of speciation (Higashi *et al*. 1999; M'Gonigle *et al*. 2012). Speciation is thus complex and models of speciation seek to identify the important processes and their dynamics, in order to provide a testable framework (Gavrilets 2003).

In recent years there has been shift in speciation research away from traditional geographic models towards assessing the forms of selection on adaptive and reproductive traits and their relationship with gene flow. Traditional geographic models go as far back as Mayr (1942) who thought that extrinsic geographical barriers to gene flow were important factors in speciation. These models include allopatric speciation where speciation occurs in geographic isolation, i.e. without gene flow; sympatric speciation where populations are panmictic or overlap in their geography; and parapatric speciation, an intermediate of the previous two in which geographic

separation occurs but is imperfect, allowing some level of gene flow. However, recent models show that although geographic barriers are unquestionably important during speciation, divergence can occur in sympatry or allopatry highlighting that geography may tells us little about the process of speciation itself (Gavrilets 2003; Van Doorn 2004). The shifting view of speciation has occurred through a dissection of the speciation process into interacting component parts that include; modes of selection (e.g. environmental and/or sexual, disruptive), demographic or intrinsic biological aspects (population size, mutation, recombination and gene flow) and the genetic/genomic basis of speciation (e.g. genetic architecture of adaptive traits, gene expression variation, the role of chromosomal inversions, neutral versus selected genomic regions and 'islands'/'continents' of genetic divergence). Thus to understand how speciation proceeds we must identify general rules and patterns for the dynamics of population divergence (Gavrilets 2003).

One example of the shifting emphasis in speciation research is an increasing interest in sympatric speciation. Sympatric speciation has remained a controversial subject for many years (Mank 2009) due to the role of gene flow; how can species differ whilst gene flow and recombination continue to homogenise the genome, preventing divergence? However, theoretical work has demonstrated that speciation with gene flow is more likely to occur with strong disruptive selection (whether this is ecological, sexual or a combination), a form of isolating mechanism causing assortative mating, strong association between genes influencing fitness and genes causing non-random mating, high genetic variation and minimal constraint on being choosy in mating decisions (Fitzpatrick *et al*. 2009; Gavrilets 2003; Gavrilets & Hayashi 2005; Kirkpatrick & Ravigne 2002). Whether speciation with gene flow occurs regularly in nature remains to be seen, with some arguing that the conditions

needed are too specific (Gavrilets 2005) and others that it might occur easily (Doebeli *et al*. 2005).

Ecological speciation

The shifting views on speciation in the last 12 years have included a renewed interest in the role of ecology with a focus on ecological speciation. It is within this ecological framework that this thesis examines population divergence. Ecological speciation occurs through the adaptation of populations to different environments, with reproductive isolation evolving directly or indirectly through increasing incompatibility of genetic regions (Nosil *et al*. 2009b; Rundle & Nosil 2005; Schluter 2000, 2001). Exactly how ecological speciation differs from other modes of speciation is subtle, mainly because speciation in most cases will involve an organism's ecology (Sobel *et al*. 2010). For example mutation order speciation involves environmental selection, but the source of variation occurs through differing mutations that arise across different populations that are experiencing the same ecological pressures (Schluter 2009). Conversely, ecological speciation occurs through divergent selection regimes across different ecological settings. The environment (broader sense ecology) includes the natural environment or niche, as well as interactions with other individuals, e.g. frequency-dependent interactions through competition and predation. Ecological speciation arises through the selective power of the environment but does not include selective pressures derived from the interactions between the sexes, or sexual selection (although it does include those cases where sexual selection and ecological selection are linked e.g. when ecological selection drives divergence in reproductive characters/preferences; Boughman 2002; Rundle & Schluter 2004; Schluter 2000). It also does not include cases where

speciation occurs through genetic drift, although it is thought that speciation under genetic drift is less likely to occur (Coyne & Orr 2004; Turelli *et al*. 2001). Additionally, ecological speciation can occur regardless of geographical situation, and includes cases of both the incidental evolution of reproductive isolation and direct selection for reproductive isolation. Although the ecological environment does not depend strictly on geographic distance, allopatry and sympatry are included in this new emphasis, and are thought of in terms of gene flow and thus as a more continuous parameter.

Indirect ecological speciation involves the evolution of reproductive isolation as a by-product of environmental adaptation (Rundle & Schluter 2004). This process may occur with gene flow, and involves divergent selection between environments that leads to phenotypic changes, some of which might be involved in reproduction. If adaptive changes in behaviour, morphology or physiology also cause assortative mating as a by-product (so-called 'magic' traits; Servedio *et al*. 2011), this might lead to reproductive isolation. Sexual selection might strengthen this isolation, if changes due to adaptation occur for sexually selected traits (i.e. shifts in male reproductive traits and female preferences). Direct selection for reproductive isolation during ecological speciation involves the reduced fitness (viability/fertility) of offspring produced from heterospecific matings, or reinforcement (Dobzhansky 1940). This can occur in the presence of gene flow, where offspring have reduced fitness in either environment leading to selection for discrimination of heterospecifics (evolution of mating preferences to recognise locally adapted conspecifics) and premating isolation. It can also occur upon secondary contact of previously allopatric species, where ecological incompatibilities may have already arisen. It should be noted that fitness costs to mating heterospecifically during reinforcement may actually be borne by the

mating individual, for example through courtship costs or predation risks, rather than just the hybrid offspring (Rundle & Schluter 2004; Servedio 2004).

The genetic basis of ecological speciation is increasingly being addressed at the genomic level, with models such as divergence hitchhiking, which lead to genomic 'islands' and 'continents' of divergence, coming to the fore (Feder & Nosil 2010). These models come from the idea that genomic divergence during speciation is heterogeneous, with some regions being genetically divergent and some being homogenised by gene flow or genetic drift (Smadja *et al*. 2008; Turner *et al*. 2005; Via & West 2008). Speciation is likely to start as a slow process with restricted gene flow at only a few key loci (or genomic islands) involved in local adaptation, sexual conflict, mate choice or other selective processes (Butlin 2010). These divergent loci then slowly spread across the genome, due to genetic linkage and reducing gene flow and interpopulation recombination, until each genome is a single coadapted unit, or genomic continent (Charlesworth *et al*. 1997; Feder & Nosil 2010). This genic view of speciation has been discussed previously (Noor & Feder 2006; Via 2001; Wu 2001), and has been suggested as a potential mechanism for sympatric speciation. However, theoretical models show that the spread of divergent genomic regions requires specific conditions, such as small effective population sizes, low gene flow and strong selection on multiple loci (Feder & Nosil 2010). Chromosomal inversions (where incompatible loci have accumulated) would aid in the spread of islands (Noor *et al*. 2001), and sexual selection and pleiotropy may provide the strength of selection needed for spreading to occur (Feder & Nosil 2010; Wolf *et al*. 2010). However, the process of divergence hitchhiking remains poorly understood and empirical evidence for the process itself is lacking (Berlocher & Feder 2002; Via 2001).

There are now an increasing number of studies that identify 'outlier' loci, which are assumed to be a result of divergent selection (see Nosil *et al*. 2009a for some examples). The idea of population genomics (Lewontin & Krakauer 1973; Luikart *et al*. 2003), measured through such genomic scans, is to quantify genetic divergence using a common measure such as F_{ST} (Hudson *et al*. 1992; Wright 1949) and identify genetic regions that are selectively neutral, as well as those that are divergent due to the action of selection. Several statistical methods can be used to determine whether a locus behaves as an outlier, including F_{ST} tests against a neutral expectation derived from either the genome-wide empirical distribution (Akey *et al*. 2002) or neutral coalescent simulations, which might use parameters estimated from the data (Stinchcombe & Hoekstra 2008). This is a potentially powerful system because, for two diverging populations, demographic properties can be assessed using methods that assume neutrality, such as population size and gene flow, but divergent regions can also be assessed in terms of their selective role in adaptation and reproductive isolation. New methods for analysis of high throughput data can identify and partition these different data types (e.g. Jones *et al*. 2012) and such technologies now permit identification of complete sets of adaptive loci through statistical associations between genetic variation and phenotypes (e.g. traditional QTL studies and GWAS studies). However, there are several problems in identifying outlier loci, linked to the limitations of the method (Butlin 2010). Not only do you need full genome techniques for the method to capture all the potential loci under selection, but you also need the full genome of the organism, with good coverage for reliable sequence verification, to map significant loci to functional genomic regions (i.e. adaptive genes). Such methods also require a representative sample of the populations you are comparing. Further, methods for detecting divergence such as F_{ST} can be

sensitive to demographic factors such as bottlenecks, potentially obscuring the neutral-selective division of loci. However, decreasing expense of high-throughput sequencing and improvement in both neutral models and methods for detecting selection will aid in the characterisation of divergent loci.

Speciation genes

To understand how speciation occurs it is necessary to investigate how it proceeds from the earliest stage of reproductive isolation, and identify the phenotypes involved. Investigating the genetic architecture of these phenotypes will uncover key genes involved in reproductive isolation, or 'speciation' genes (Noor 2003). Much work has focussed on postzygotic speciation genes because effects such as hybrid male sterility are often the first obvious signs of speciation (Orr *et al*. 2004; Sun *et al*. 2002). As such speciation genes could be defined as loci for which the allelic form of one population is not compatible with the genome of another (Noor 2003; Ritchie & Noor 2004; Sun *et al*. 2002). However, a broader definition of speciation genes is any gene that contributes to reproductive isolation between populations, including both postzygotic and premating phenotypes, regardless of whether it caused the initial isolation or not (Nosil & Schluter 2011; Wu & Ting 2004).

The genetic architecture of traits underlying reproductive isolation is still largely unknown (Mackay *et al*. 2005). In fact there is little recent literature with a broad and general treatment of the genetic mechanisms underlying premating isolation in particular (Etges & Tripodi 2008; Ritchie & Phillips 1998). Yet characterising the genetic sources of important phenotypic variance could provide valuable and unique insights into the processes driving divergence. One reason for this paucity of knowledge is the complexity of the genetic architecture of speciation,

which calls into question the simplistic idea of a speciation gene (Takahashi & Ting 2004). Most behavioural traits are polygenic, with many different factors influencing the final phenotype. The genetic architecture of a quantitative trait includes: the number of loci and alleles, the genomic distribution of loci, the magnitude of effects of alleles, the directionality of alleles, allelic relationships (dominance and epistasis), pleiotropy and the ploidy and mode of inheritance (Shaw & Parsons 2002). Two extreme types of architecture have been identified that illustrate the range of genetic basis in which phenotypes may lie. Type I is a model of many genes with small, additive effects; type II involves major or modifier genes, of strong effect, underlying the phenotype in question (Gleason & Ritchie 2004; Templeton 1981). These types are at extremes and the importance of each is still under debate.

The search for speciation genes should be more accurately described as the search for the molecular basis of speciation (Wolf *et al*. 2010). This concept includes our expanding knowledge of the molecular basis of phenotypes, and is an integrative examination (often called systems genetics) of not only the genetic architecture, but also how these genes are regulated and coexpressed as gene networks to form the phenotypes involved in reproductive isolation (Mackay *et al*. 2009; Springer *et al*. 2011). This includes mechanisms of expression regulation as well as regulatory loci (whether in *cis* or *trans* to the gene it regulates), both DNA and RNA based. Many identified causative SNPs now being discovered substantial distances from known gene encoding regions and presumed to be transcription factor binding sites or other *cis* regulatory features (Frazer *et al*. 2009). However, identifying the effect of a causative SNP on a phenotype in a non-coding region is challenging without more knowledge on the functional mechanics of such regions. For example many new types

of non-coding short and long RNAs have recently been identified that have complex roles in gene expression (Mercer *et al*. 2009).

Further knowledge is also required on the types and functions of genes that underlie reproductive isolation. Although several examples of speciation genes have been uncovered (see Coyne & Orr 2004 for some examples; Orr *et al*. 2004; Rieseberg & Blackman 2010; Rundle & Nosil 2005; Swanson & Vacquier 2002) there are still surprisingly few, particularly for premating phenotypes. Functions of these loci range from broad roles such as nucleoporin genes or transcription factors, to more specific proteins such as *YUP* in monkeyflowers, which influences flower colour and pollinator isolation (Nosil & Schluter 2011). The function of speciation genes is often likely to depend on the particular species under study, and the search for speciation genes (as for genes underlying adaptive phenotypes or causing disease) is challenging. Often several lines of study will be needed simply to identify a candidate gene, with further work necessary to confirm the involvement of that locus. The strongest of such confirmations includes experimental manipulation of a candidate gene through transgenics or gene knockouts, coupled with expression and phenotype analyses (Nosil & Schluter 2011).

An important step towards understanding speciation is to identify the processes that cause the genetic divergence. However, determining the processes involved, for example detecting signals of selection or drift, still present problems. Examples of these issues include how to detect selection in non-coding, regulatory DNA and how to detect selection at a genome-wide level (Nielsen 2005; Sabeti *et al*. 2007). Such methodological challenges are currently being addressed using new techniques such as genome-wide association studies and genome-wide scans to identify large numbers of loci associated with phenotypic differences and genetic

divergence (McCarthy *et al*. 2008; Oleksyk *et al*. 2010). Further, expression QTL and quantitative trait transcripts (eQTLs and QTTs) analyses can be used to detect genomic regions through correlating gene expression to phenotypes, similar to the traditional genetic-based QTL analyses (Mackay *et al*. 2009). With an increasing number of genomic technologies and genomes becoming available, the intersection of multiple data types is now possible, but this also presents new challenges in terms of large-scale data analysis, statistical intersection of different data types and interpretation of results.

Determining the molecular basis of reproductive isolation in incipient species is the first step to understanding how speciation occurs, and the mechanisms that underlie it. However, other than the knowledge that phenotypes causing isolation are commonly polygenic, there is still a lack of data and general rules concerning the underlying genetics of behavioural isolation (Arbuthnott 2009). Due to this polygenic control, many forces may influence courtship behaviours. The environment in which an organism lives will create selection pressure as well as influence how genes are expressed. Polygenic courtship traits often demonstrate phenotypic plasticity that may affect mate choice and sexual selection across environments, and little is known of the role of plasticity generally in speciation (Butlin *et al*. 2012). Plasticity may drive or inhibit population divergence, and if plasticity and genotype-by-environment interactions play a role in the speciation process, this may also obscure the discovery of speciation genes.

Box 1.1: Overview of terms and concepts associated with phenotypic plasticity

Phenotypic plasticity (Schlichting & Smith 2002)
Any change in an organism's phenotype in response to an environmental signal.

Genotype-by-environment interactions (Pigliucci 2001)
Genetic variation in phenotypic plasticity that leads to non-parallel or overlapping reaction norms between genotypes.

Phenotypic accommodation (West-Eberhard 2005b)
The adjustment, without genetic change, of aspects of a phenotype following a novel input during development.

Genetic accommodation (Baldwin 1896, 1902; West-Eberhard 2005a)
Heritable variation in plastic response that can selected upon once the phenotype is induced. Selection may change the mean of the trait but not the level of plasticity, or act to change the plasticity of that trait as well.

Genetic assimilation (Pigliucci *et al.* 2006; Waddington 1942)
A novel or adaptive phenotype is induced and fixed through selection whereby the induced phenotype no longer requires the input, reducing the plasticity of, or canalizing, the trait.

Genetic compensation (Grether 2005)
When a plastic response is maladaptive, selection acts in a direction opposite to the response so that genetic change can compensate.

Phenotypic modulation (Smith-Gill 1983)
Non-specific phenotypic variation results from environmental influences on rates or degrees of the developmental program.

Developmental conversion (Smith-Gill 1983)
Organisms use specific environmental cues to activate alternative genetic programs controlling development.

Allelic sensitivity (Schlichting & Pigliucci 1995)
The product of different alleles might be 'sensitive' to environmental changes in a different but consistent fashion.

Regulatory plasticity (Schlichting & Pigliucci 1995)
Differential response in regulation of a pathway of gene expression through specific regulatory genes.

Canalization
A developmental phenomenon in which a trait is robust to changing environmental conditions.

Broad dimensional aspects of phenotypic plasticity
> *Phenotypic category of response*: Physiological/Behavioural/Morphological.

> *Hierarchical level of response*: Gene expression/Cellular/Tissue/Phenotype.

> *Source of plastic response*: Mutational change/Internal environment/External environment.

> *Temporal dimensions*: Window of developmental opportunity/Timing and rate of plastic response/spatiotemporal environmental heterogeneity.

> *Adaptive versus non-adaptive response*

Phenotypic plasticity and genotype-by-environment interactions

Phenotypic plasticity provides a way for a single genotype to produce different phenotypes across different environments. Phenotypic plasticity has previously been viewed as an obstacle to population divergence, yet recent evidence suggests it could have an important role in ecological speciation (Butlin *et al*. 2012; Pfennig *et al*. 2010). In a review of phenotypic plasticity in plants, Bradshaw (1965) set out several outstanding questions on plasticity; What is the mechanistic basis of continuous and discreet plasticity? How are the plastic responses of different traits related? What is the genetic control (or molecular architecture) of plasticity and can it be selected? How much genetic variability for plasticity exists in natural populations? Schlichting (1986) noted that though we have gained some understanding of plasticity, these essential questions remain unanswered.

The investigation of plasticity from different viewpoints has led to a set of overlapping definitions of related processes (Box 1.1). These have been classed in terms of development (e.g. phenotypic modulation and developmental conversion) and genetic (e.g. allelic sensitivity and regulatory plasticity) processes. However, these terms are inextricably linked as for example, allelic sensitivity is likely to result in phenotypic modulation throughout development (see Pigliucci 2001 for a thorough dissection). Pinning down a single definition of phenotypic plasticity is a challenge due to the differing interpretations of what a plastic trait is. A useful, if broad, definition was proposed by Schlicting and Smith (2002) that 'phenotypic plasticity is any change in an organism's characteristics in response to an environmental signal'. As Schlicting and Smith (2002) state, this definition encompasses all hierarchical levels of an organismal plastic response (from gene expression, gene networks or a cellular response all the way up to the phenotype, see Box 1.1) as well as not

assuming any particular phenotypic category of response (behavioural, physiological or morphological) or environmental input (internal, external or new mutation in different genetic backgrounds).

Phenotypic plasticity is often measured using reaction norms, which chart trait changes across continuous or discrete environments (Fig. 1.1). Individual reaction norms represent a single genotype, with multiple genotypes each having reaction norms at a population level, representing population genetic variation for a particular phenotype (Fig. 1.1a). If these reaction norms are parallel then the level of plasticity across environments is equivalent for all genotypes and thus there is no genetic variation for a plastic response (Fig. 1.1b). However, if the reaction norms of two genotypes are not parallel this represents genetic variation for plasticity, or a genotype-by-environment interaction (GEI, Fig. 1.1c). An extreme case of GEI is when two reaction norms cross, often called ecological crossover (Fig. 1.1d, Greenfield & Rodriguez 2004).

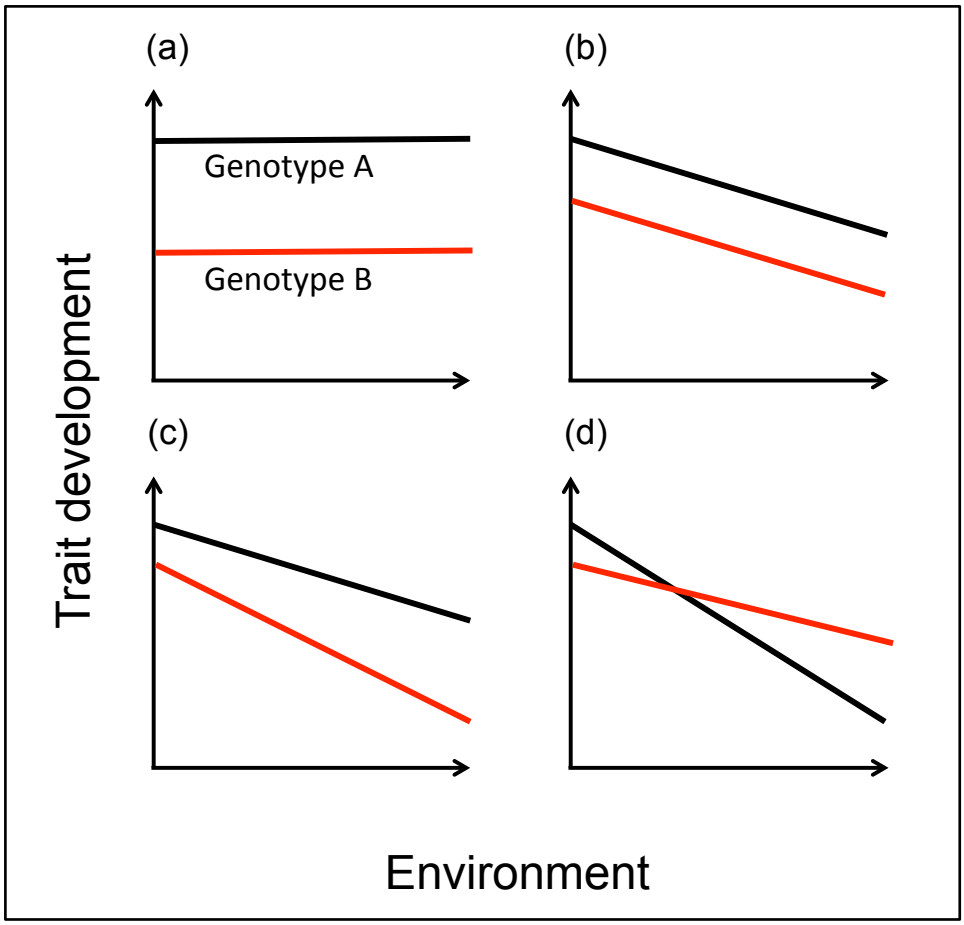


Fig. 1.1 Example of reaction norms of two genotypes within a population. (a) two genotypes, demonstrating genetic variation for a trait but no plasticity or genetic variation for plasticity. (b) Both genotypes are plastic, but there is no genetic variation for plasticity. (c) Both genotypes are plastic to differing degrees (non-parallel lines) thus demonstrating genetic variation for plasticity or GEI. (d) Crossing reaction norms indicate GEI often termed ecological crossover.

Many of the terms listed in Box 1.1 attempt to define the evolutionary importance of plasticity. Selection can act on the mean of a population trait distribution, as well as on the variance of distributions across environments. Assuming there is genetic variation for plasticity, environmental heterogeneity and that the plasticity of a particular trait increases the fitness of an organism, there will selection for plastic genotypes (Schlichting & Pigliucci 1995). Selection can lead to a change in the

population mean of a trait in an environment, but involve little change in the level of genetic variation for plasticity itself, or it can influence both (genetic accommodation; Box 1.1). Figure 1.2 summarizes some potential effects of selection on mean and variance of trait values after a phenotype is induced. Thus, selection can act to shape reaction norms, selecting for optimal plastic genotypes in a population and potentially eroding genetic variation for plasticity. However, selection may act to reduce the plasticity of a trait, leading to the assimilation of a previously plastic response genetically (genetic assimilation; Box 1.1; Fig. 1.3). Genetic assimilation leads to canalization, where the expression of a trait remains stable regardless of environmental and genetic change, and results in the fixation of a phenotype even when the environmental input is removed (Crispo 2007). Plasticity and GEIs are important in the production of many phenotypic characters and are potentially important during ecological speciation where populations experience different environments.

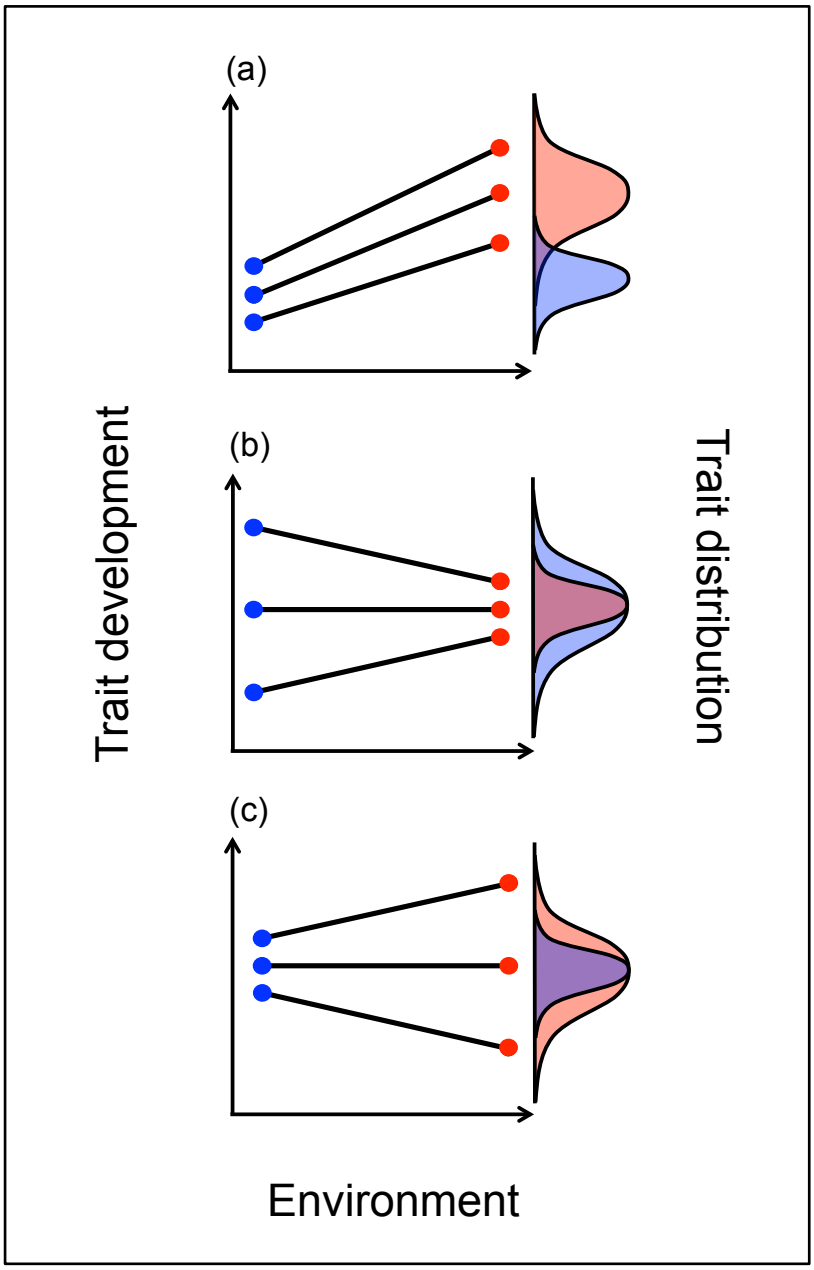


Fig. 1.2 Potential scenarios for plastic phenotypic response across environments, which may have been produced through selection and can be shaped by selection. (a) An increase in the mean and variance of a trait across environments. (b) A decrease in the variance of the trait across environments, mean unchanged. (c) An increase in the variance of the trait across environments, mean unchanged. Modified from Fordyce (2006).

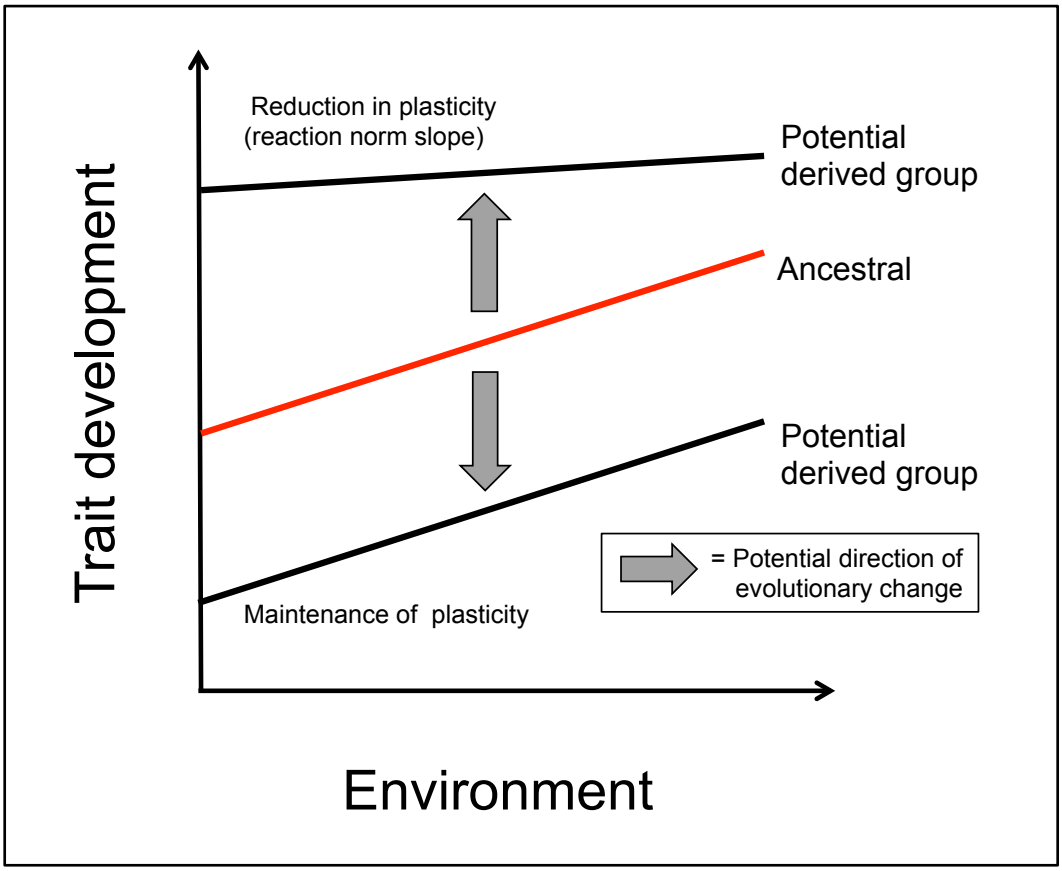


Fig. 1.3 Effect of selection on plasticity within a population. Each reaction norm represents the average plasticity of a population of genotypes. The up arrow demonstrates selection for less plastic genotypes, decreasing a population's phenotypic plasticity (genetic assimilation). The down arrow represents the maintenance of phenotypic plasticity. Both trait means move away from the ancestral trait mean.

Plasticity and ecological speciation

Phenotypic plasticity plays an important role in ecological speciation because it allows organisms to respond adaptively to environmental variability, aiding ecological selection in driving rapid population divergence. However, phenotypic plasticity has previously been seen as an alternative to ecological adaptation (the first step in ecological speciation) rather than facilitating it (Wright 1931), though this view was not held by some early proponents of plasticity (e.g. Waddington 1953). Plasticity allows an adaptive response to new environments, allowing the colonization of even very divergent environments. Evidence shows that plasticity can aid in persistence of species in new environments, or contribute to reproductive isolation and genetic divergence between environments (Price *et al*. 2003; Thibert-Plante & Hendry 2011; West-Eberhard 1989; West-Eberhard 2003; West-Eberhard 2005a; Wund *et al*. 2008). Phenotypic plasticity has been reviewed extensively (see Fitzpatrick 2012; Pfennig *et al*. 2010 for recent reviews), and a number of theoretical models have examined plasticity across both spatially and temporally varying environments (see Thibert-Plante & Hendry 2011 for a succinct summary).

Phenotypic plasticity is likely influence ecological speciation through its two main components: ecological adaptation and reproductive isolation. Plasticity can occur for ecological or reproductive traits, however the evolutionary result (genetic divergence between environments, continuation in an environment, or extinction) is complex and will depend on several linked aspects of the developmental response. First, the extent to which any response (phenotype) is optimal (adaptive) for the environment in which it occurs (including suboptimal and nonadaptive responses) dictates levels of divergent selection. This optimality may depend on the second aspect, the constraint (costs and limitations) of a response, which is linked to how

different a newly colonized environment is, and the presence of plasticity in functionally important traits. Finally, the evolution of plasticity is dependent on the spatial/temporal aspects, including environmental heterogeneity (or gene flow) and developmental sensitivity (time during development in which a plastic response can occur). Adaptive phenotypic plasticity has received more attention in the past and thus each of these aspects have been more thoroughly studied in an adaptive rather than nonadaptive context.

Phenotypic plasticity might often be adaptive, i.e. one that increases an organisms fitness in a given environment (DeWitt & Scheiner 2004; Thibert-Plante & Hendry 2011). If a plastic response is near optimal in a new environment this will lead to persistence and little or no divergent selection (DeWitt & Scheiner 2004). However, suboptimal plasticity in one environment means that divergent selection may occur across environments (Thibert-Plante & Hendry 2011). A plastic response might be suboptimal for a number of reasons including certain constraints, costs and limitations, to the developmental response. The costs of plasticity are the fitness deficits associated with plastic genotypes relative to fixed genotypes producing the same mean phenotype in the focal environment, whereas limitations are the functional constraints of a plastic response to an environment, producing a less optimal or non-adaptive response (DeWitt 1998; DeWitt *et al*. 1998). These costs and limits will influence the evolution of plasticity and if adaptive plasticity is less than optimal, divergent selection can potentially exist.

Important to the adaptive role of phenotypic plasticity is the spatiotemporal heterogeneity of the environment, which is crucial to whether populations maintain genetic variation for plasticity or if this variation is eroded. Environmental variation may be temporal, or gene flow between environments. Such variation is called the

'grain size' of an environment, and is a measure of environmental change within the lifetime of an individual (Levins 1968). In coarse-grained environments individuals experience a single environment whereas in fine-grained environments they experience a succession of different environments (Gillespie 1974). Fine-grained environments impose stronger selection for an adaptive plastic response (Baythavong 2011; Hollander 2008; Lande 2009), depleting genetic variation for plasticity. With decreasing environmental heterogeneity the strength of selection for a particular plastic response decreases, the extreme of this being a stable environment, where populations adapt to stable or prevalent conditions often reducing plasticity due to its increased cost. Developmental sensitivity is key to the evolution of plasticity and interacts with environmental grain size. For example, fine-grained environmental change after the period of sensitivity has little effect, and the environment is then effectively coarse-grained.

A recent study by Thibert-Plante and Hendry (2011) used individual-based simulations to explore the different aspects of phenotypic plasticity that influence the evolution of reproductive barriers during ecological speciation. They showed that adaptive plasticity evolves readily in the presence of dispersal between populations in different environments, and encourages the colonization of highly divergent novel environments. Thus without plasticity new environments may never be colonized, once they are and populations can persist, they may be under divergent selection (potentially on separate traits to those allowing the initial persistence). Plasticity might also be important in the evolution of reproductive isolation if plastic phenotypes are involved in reproductive success (e.g. a courtship trait/preference or pleiotropic links between ecological and reproductive characters), or if natural selection acts against hybrids and migrants with reduced fitness.

A strong link to fitness can be seen when reaction norms cross (ecological crossover; Fig. 1.1d) and fitness rankings of genotypes across environments become reversed. If the traits involved are female preference and/or male traits under sexual selection, this might lead to population divergence and reproductive isolation (depending on levels of gene flow and in which environment plasticity occurs; Greenfield & Rodriguez 2004). If environmental change is frequent, covariation between male trait and female preference might decrease if costs of plasticity are high (i.e. male signals are unreliable), yet this is balanced by increased genetic variation due to high gene flow, which would increase the benefits to a female of being more selective during mate choice. However, if environmental change becomes rare (gene flow reduces), there might be a build up of differences between populations in male signals and female preferences, causing divergence and reproductive isolation (Greenfield & Rodriguez 2004). Alternatively, reproductive isolation might evolve through selection against hybrids from different environments. For example polyphenism is the production of divergent, environmentally triggered resource-related phenotypes and can lead to spatially segregated morphs or ecotypes that display low levels of reproductive isolation (Pfennig *et al*. 2010). Once these ecotypes emerge, hybrids with reduced fitness will be selected against, favouring the evolution of reproductive isolation (Rundle & Nosil 2005).

Thibert-Plante and Hendry (2011) also demonstrated that natural selection against migrants might actually be more important in the formation of reproductive barriers during ecological speciation than sexual selection against migrants, and selection against hybrids (i.e. selection against migrant offspring). This is because natural selection often acts earlier than sexual selection, following the dispersal of migrants to a new environment, and before migrants produce offspring (Hendry 2004;

Thibert-Plante & Hendry 2011). They showed that reproductive isolation was weaker when plasticity is expressed after dispersal as migrants are better suited to the environment in which they developed. Conversely, reproductive barriers were stronger (or unaffected) when plasticity is expressed after dispersal and selection can act. These results highlight the importance of the timing of dispersal in relation to the developmental timing of a plastic response (developmental sensitivity). Dispersal after a plastic response will lead to selection against sub-optimal migrants in the new environment and thus is a reproductive barrier. For example, change in the timing and rate of phenotypic development (heterochrony) can be induced by the environment and might be particularly important for the production of ecotypes (Parsons *et al*. 2011). It should be noted that such plasticity causes only weak reproductive barriers, yet these barriers occur *before* genetic divergence due to selection against phenotypes induced in the alternative environment. Thibert-Plante and Hendry (2011) note that this inverts the causal pathway to reproductive isolation, which is commonly thought to begin with a certain level of adaptive genetic divergence. This might be very important at the initial stages of ecological speciation by reducing gene flow and allowing adaptive divergence.

The environment plays a joint role in shaping phenotypes in that it acts to induce phenotypic variation through plasticity (developmental) as well as to shape phenotypes through selection during adaptation (transgenerational). The interaction between environmental plasticity and environmental selection can lead to change in levels of adaptive plasticity, which can influence adaptation and speciation. Non-adaptive plastic responses can also play a role in evolutionary change if non-adaptive plasticity in a new environment creates an increase in trait variance, releasing cryptic

variation upon which selection can act (Ghalambor *et al*. 2007; Le Rouzic & Carlborg 2008).

Adaptation to different environments might occur through an invasion of a novel environment with continued gene flow. Alternatively gene flow might be reduced due to either intrinsic or extrinsic factors, strengthening divergent selection across environments. Thibert-Plante and Hendry (2011) demonstrated that phenotypic plasticity allows for the colonization of even divergent environments. Once colonization has occurred, plasticity can enhance or reduce reproductive isolation, depending crucially on the timing of a plastic response compared to dispersal time during development. This has implications for the types of plastic traits that contribute to reproductive isolation as those that do are likely to be responsive to the environment early in development, before dispersal. The colonization of new environments allows for further phenotypic change, which may or may not involve plasticity, but certainly would not have occurred without it (Fitzpatrick 2012). Phenotypes derived from a plastic response can be shaped by genetic accommodation and assimilation (Lande 2009), potentially fixing divergent phenotypes across environments. Therefore it is clear that phenotypic plasticity plays a role in encouraging the colonization of new environments, can buffer species existence, and can play a role in the evolution of reproductive isolation.

The molecular basis of phenotypic plasticity

Understanding the molecular basis of phenotypic plasticity will aid in our understanding of the genetic basis of adaptation and speciation. Genetic variation for plasticity has been demonstrated in several studies and is thought to be widespread in nature (Levine *et al*. 2011; Pigliucci 2001). Yet little is known about where this

variation lies (e.g. genomic locations of regulatory regions) or the functions of genes such variation influences. The latter is addressed in Chapter 4 where I examine the functions of genes that show plastic expression variation according to ecological environment. An adaptive plastic response is likely to involve molecular pathways and loci that have evolved to allow a plastic change in accordance to an environmental input. Such 'plasticity genes' as proposed in Schlicting and Pigliucci (1993) can be defined as 'regulatory loci that directly respond to a specific environmental stimulus by triggering a specific series of morphogenic changes' (Pigllucci 1996). This definition refers to but, as noted by Pigliucci (2001), is not limited to genes regulating morphology, and behavioral and physiological changes should also be included. Aubin-Horth & Renn (2009) note that 'phenotypic plasticity can be defined as a re-programming of the genome in response to the environment' referring to mechanisms of gene expression regulation. However, it should be noted that not all regulatory genes (and not all plastic gene expression) necessarily contribute to an adaptive plastic response (Pigliucci, 2001).

The molecular basis of plasticity is likely to be complex due to the multiple mechanisms controlling gene expression. For example Levine *et al*. (2011) discovered plasticity in gene expression in a genotype-by-environment interaction across tropical and temperate *D. melanogaster* populations, according to temperature. They linked expression variation to chromatin based regulatory control, which coincided with a genetic signature of strong selection in chromatin remodeling factors in the temperate population. Thus chromatin-based control of gene expression was shown to be involved in adaptive evolution and may be a common mechanism underlying plasticity. Genetic variation in 'plasticity genes' may consist of variation in a diverse set of regulatory features. For example transcription factor binding sites changes

influence gene expression and rapidly evolve, and promoter region changes have been linked to chromatin-based regulation of gene expression (Tirosh *et al*. 2008). The recent discovery that epigenetic modifications to DNA, RNA and proteins influences gene expression suggests that epigenetic mechanisms might also play a role in plasticity and the production of quantitative trait variation (Johannes *et al*. 2009; Johnson & Tricker 2010; Richards 2009), and I provide further evidence for this in Chapter 4.

The combining of ecology, evolution, development and molecular and genetic viewpoints (recently described as integrative biology) might help to uncover the molecular mechanisms underlying plasticity (Aubin-Horth & Renn 2009). The increasing use of modern molecular techniques such as microarrays, high throughput (next generation) sequencing and ChiP-seq allows the examination of not only gene expression, but mechanisms that increase protein diversity such as alternative splicing and controls of gene expression such as non-coding RNAs (ncRNAs), DNA methylation, chromatin remodeling and other protein modifications. Using these techniques we can now examine the genome, proteome and methylome at high resolution and potentially move towards a greater understanding of how the expression of genomes works. Because we can probe at the single cell level we can compare cell types and tissue types and build an understanding of how gene expression varies over time and space. This increasing use of technology can thus allow us to start to tease apart the various ways in which phenotypic plasticity is viewed (Cossins *et al*. 2006).

Drosophila courtship behaviour and incipient species

Drosophila melanogaster has been a model organism in evolutionary studies for many years along with many other *Drosophila* species. For example, Morgan and Bridges (1916) used *D. melanogaster* to investigate heredity and Dobzhansky (1948) studied *D. pseudoobscura* and *D. persimilis* to investigate aspects of speciation and selection. Their ease of handling in the laboratory and short generation time makes them ideal for studying evolution and speciation, at least from a genetic perspective. However, little is known about their ecology, making them less ideal for the study of ecological speciation. Much is known about the genetics of *D. melanogaster* and the full genome was published in 2000 (Adams *et al*. 2000), and since then 11 other *Drosophila* species have had their full genomes sequenced (*Drosophila* 12 Genomes Consortium 2007).

In *Drosophila* courtship behaviour is usually initiated by males and involves visual, acoustic, olfactory, gustatory and tactile cues for female evaluation of potential mates (Greenspan & Ferveur 2000). There is great variation in courtship behaviour between species and the following represent the major steps common to most species. Males begin by orienting towards a female, slightly raising their bodies. They then tap their foreleg against the female, possibly transferring contact epicuticular hydrocarbons in the process. Next, males move their wings at different angles, vibrating them to produce a species-specific courtship song. After singing, the male may circle the female and vibrate his forelegs against the abdomen of the female, extend his proboscis and lick the female genitalia, finally mounting her (Hall 1994). The exact content and order of these courtship signals may depend on the species, for example Hawaiian species of *Drosophila* have modified labellar tools which grasp female genitalia (Spieth 1974) and signals may vary between species in importance

for mate recognition (Ferveur 2005). Sex and reproductive related genes are thought to evolve rapidly, making premating courtship signals ideal candidate phenotypes for investigating the initial stages of reproductive isolation (Gleason & Ritchie 1998; Templeton 1981).

Much has been learnt from investigating incipient species. For example Dobzhansky and Pavlovsky (1967) investigated incipient species within *D. paulistorum* in which there are at least five distinct species, several being reproductively isolated with varying strengths of sexual isolation. From this species complex Dobzhansky and Pavlovsky (1966) found a strain of *D. paulistorum* that had evolved in the laboratory over a number of years to become a separate species. Other examples of closely related species aiding speciation research include *D. pseudoobscura* and *D. persimilis*. Dobzhansky (1936) examined races of *D. pseudoobscura* to localize sterility factors in hybrids and demonstrated regions of chromosomes that were responsible for sterility, or postzygotic isolation. Machado *et al*. (2002) analysed the divergence of *D. pseudoobscura* from its close relatives *D. pseudoobscura bogotana* and *D. persimilis*, finding evidence for speciation with gene flow, with loci associated with reproductive isolation showing little or no gene flow.

One interesting example of incipient speciation comes from *D. melanogaster*. The Zimbabwe *D. melanogaster* population demonstrates unidirectional sexual isolation from the more globally distributed cosmopolitan race. Zimbabwe females prefer males from Zimbabwe whereas cosmopolitan *D. melanogaster* females have only a weak preference for their own males (Wu *et al*. 1995). There is a lack of postmating isolation between the two populations, however they are divergent in premating behaviour, and these populations are considered incipient species (Hollocher *et al*. 1997a; Hollocher *et al*. 1997b). Further divergent populations of

Drosophila melanogaster have been described from West Africa, the Caribbean and the United States (US). The US population demonstrates partial sexual isolation from Caribbean populations and the African races (Yukilevich & True 2008). Evidence suggests a role for pheromone variation in the sexual isolation of these populations, and studies have identified a gene involved in courtship pheromone production called *desaturase-2* (*desat2*), which is responsible for differences in epicuticular hydrocarbon (used as contact courtship pheromones in *Drosophila*) composition in flies (Coyne *et al*. 1999; Dallerac *et al*. 2000; Takahashi *et al*. 2001). It has been shown that Caribbean populations are divergent from US populations in *desat2* allele frequencies, male morphology and courtship behaviour. However, the role of *desat2* in premating isolation has been disputed (Coyne & Elwyn 2006; Grillet *et al*. 2012), highlighting the challenges associated with determining the genetic basis of complex traits.

Incipient, or newly forming, species provide a useful tool for investigating the initial stages of ecological speciation. Studying sexual isolation between incipient species (when no other form of reproductive isolation is present) is informative because such isolation is likely to be causative. Thus it is important to examine incipient species to identify key genetic changes, before further genetic differences accumulate after speciation has occurred (Wu *et al*. 1995). Such investigation is aided by knowledge on the ecology, and well-characterised behavioural phenotypes to elucidate the genetic architecture of behavioural traits (Hollocher *et al*. 1997a). Incipient species also provide an important tool because in many cases, methodologies were only possible by analysing back-crosses of closely related species, often only possible in laboratory conditions.

Drosophila mojavensis

Drosophila mojavensis presents a useful system for examining the underlying mechanisms of incipient ecological speciation because it has a well-characterized ecology. *D. mojavensis* is found in northwest Mexico and the southwestern United States (Fig. 1.4). The species is cactophilic and exists as four allopatric populations each living primarily on the necrotic tissue of a different host cactus species. *D. mojavensis* has been extensively studied and the phylogeny and ecology of the species is relatively well known (Etges & Ahrens 2001). *Drosophila arizonae* is the closest relative to *D. mojavensis* and they are thought to have diverged around 2.4 (±0.7) Mya following the formation of the Baja peninsula (Matzkin 2004; Matzkin & Eanes 2003), allowing each species to evolve in isolation. The four *D. mojavensis* populations consist of the Baja peninsula group that live and breed primarily on pitaya agria cactus (*Stenocereus gummosus*), the Mainland population that exist on organ pipe cactus (*Stenocereus thurberi*), the Mojave Desert population on barrel cactus (*Ferrocactus cylindraceous*) and the Santa Catalina Island population on prickly pear cactus including *Opuntia demissa* and *O. littoralis* (Fig. 1.4).

Evidence suggests that the Baja population is ancestral, yet the phylogeographic pattern within *D. mojavensis* remains unresolved, and is addressed in chapter 3 of this thesis. Inversion polymorphism and genetic variation at two nuclear loci indicated that the Baja population was ancestral (Johnson 1980; Matzkin 2004; Matzkin & Eanes 2003), however this has been recently challenged based on mitochondrial data (Reed *et al*. 2007) and the order and timing of population divergence is unknown. However, microsatellite genotyping indicated that these populations are distinct from one another, with little contemporary gene flow between them (Ross & Markow 2006) and the prevailing view based on ecology and genetic

information is that the Baja group is ancestral with a more recent colonization of mainland Mexico (Etges *et al*. 1999).

The colonization of mainland Mexico led to a host shift from their preferred ancestral host plant, agria, to organ pipe cactus (Etges & Ahrens 2001). Several species of cacti exist on the peninsula, yet they are rarely utilised by the Baja flies, which prefer the agria cacti as hosts. Agria cactus only exists in one small patch on the Sonoran coast (near Punta Onah, Sonora), thus it seems a shift to predominant (but not exclusive) use of organ pipe cactus has aided in the range expansion of *D. mojavensis* (Etges 1992). This host shift has caused changes in a suite of life history and reproductive traits in the Mainland population (Etges & Heed 1987). These changes include population differences in mating behaviour leading to sexual isolation from the Baja population, and the two populations are considered incipient species (Markow 1991; Pfeiler *et al*. 2009; Zouros & d'Entremont 1974).

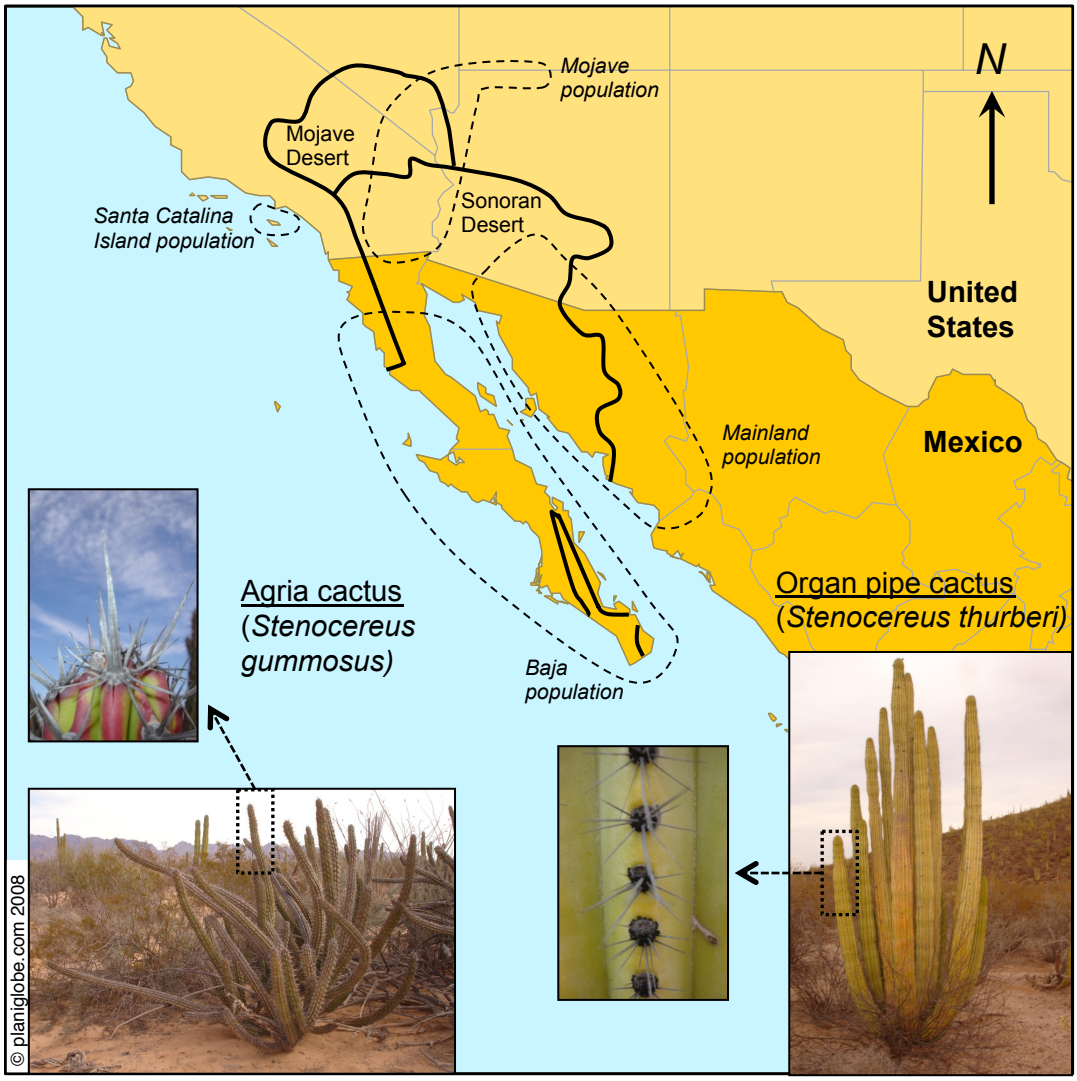


Fig. 1.4. Map of *Drosophila mojavensis* species range throughout the southwestern US and northwestern Mexico. Dark lines indicate the approximate border of the Sonoran and Mojave Deserts and dashed lines show the range of each allopatric population, with population name indicated in italics. Below are examples of the primary host cactus species for the Baja (agria cactus) and Mainland (organ pipe cactus) populations, on which the majority of this thesis focuses.

Recent publications have defined all four populations of *D. mojavensis* as subspecies (Pfeiler *et al*. 2009), however the definition of subspecies may not be a useful category in evolutionary biology due to difficulties in narrowing down an exact

definition (Wilson & Brown Jr 1953). Species definitions are contentious enough already and thus here I consistently refer to the different groups of *D. mojavensis* as populations originating from a particular geographic location e.g. 'Baja California peninsula population', or for brevity 'Baja population'. The main focus of this thesis is on the Baja and Mainland populations, which were used to investigate speciation at the incipient stages.

Unidirectional premating isolation exists between the Mainland and Baja populations of *D. mojavensis*, in which Mainland females discriminate against Baja males, and Baja females do not discriminate between males of either population (Markow 1991; Stennett & Etges 1997). Etges (Etges 1992) put forward three hypotheses concerning the mechanism for the evolution of this reproductive isolation: 1) Reproductive character displacement due to sympatry of the Mainland population with *D. arizonae*; 2) use of different host plants for feeding and oviposition causing a physiological shift in some traits; 3) changes in premating behaviour due to pleiotropic links to an adaptive fitness component associated with host plant change. There is some evidence for reinforcement between the Mainland population of *D. mojavensis* and *D. arizonae* in sympatry (Jennings & Etges 2010; Massie & Markow 2005; Wasserman & Koepfer 1977; Zouros & d'Entremont 1980), which might have led to reproductive character displacement in the *D. mojavensis* Mainland population, causing Mainland females to become more discriminate during mate choice with the Baja population. However, it is difficult to distinguish reinforcement between *D. mojavensis* and *D. arizonae* from indirect selection for reproductive characters (through ecological adaptation) in the Mainland population, both of which might have led to the evolution of premating isolation between the Mainland and Baja groups (Servedio & Noor 2003). Further, premating isolation between the Mainland and Baja

is influenced by the rearing host cactus species, suggesting a role for ecological adaptation and phenotypic plasticity (Brazner & Etges 1993; Etges *et al*. 2007; Etges *et al*. 2009). Thus, the evolution of sexual isolation between the Baja and Mainland *D. mojavensis* populations might have involved a combination of ecological adaptation to different host plants, phenotypic plasticity and reproductive character displacement between *D mojavensis* and *D. arizonae*.

D. mojavensis is considered oligophagic, typically using one host within a geographic region, yet the agria and organ pipe host plants present quite different ecological settings. Although belonging to the same genus, agria and organ pipe cacti differ in a number of ecological traits. Agria cacti have numerous, thinner stems than organ pipe cacti and grow in thick, low-standing patches, reproducing through vegetative growth (Fig. 1.4). Organ pipe cacti, however, are taller with larger stems and reproduce through fruiting and animal dispersion of seeds, growing more widely-spaced as individual plants (Fig. 1.4; Etges *et al*. 1999). Although stem size of agria cactus is smaller, the density of stems and the density of rots (necrotic tissue) is greater than that of organ pipe (Heed & Mangan 1986). Thus agria cacti represent a more continuous and ubiquitous resource, whereas organ pipe cactus rots last longer when they occur, due to larger diameter of their stems, but occur less frequently (Etges *et al*. 1999). Because the flies live on the necrotic tissue the chemical ecology of the rotting material is also important. Differences between agria and organ pipe are seen in their chemical make-up that might influence the energy level, metabolism and relative proportions of precursor fatty acids utilised by the fly (Fogleman & Danielson 2001).

Several life history and reproductive traits are population-specific and also demonstrate considerable cactus dependent phenotypic plasticity. Flies from the Baja

population express shorter egg to adult development times, higher viabilities, smaller thorax sizes, lower lifetime fecundities and slower rates of sexual maturation in the laboratory than Mainland flies, regardless of host cactus (Etges 1990; Etges & Heed 1987; Etges 1993; Etges & Klassen 1989). However, when populations are raised on each cactus, there is cactus-dependent expression of life history traits, with organ pipe generally causing longer development times, lower viabilities and smaller thorax sizes (Etges & Heed 1987), with rearing on opposing cacti causing strong genotype-by-environment interactions (Etges *et al*. 2010). Reproductive traits also demonstrate considerable geographic and host-specific variation, most notably for courtship song and contact pheromones (Etges & Ahrens 2001; Etges *et al*. 2007; Etges *et al*. 2009; Etges *et al*. 2006; Stennett & Etges 1997). Both these traits are thought to mediate the unidirectional sexual isolation between Baja and Mainland populations (Etges 2002), and the level of assortative mating is subject to genotype-by-environment interactions, depending on the rearing cactus species and the population of origin (Brazner & Etges 1993; Etges 1992, 1998).

Epicuticular hydrocarbons (CHCs) are long chain fatty acids, which in *Drosophila* act as non-volatile contact pheromones for mate recognition (Ferveur et al., 1997, Stennett and Etges, 1997, Ritchie and Noor, 2004). CHCs and courtship songs are important signals that can influence mating success in *Drosophila*, and each consist of different constituent components that can vary to produce species-specific behaviour. A set of 'diagnostic' CHC components, three alkadienes; 8,24-tritricontadiene, 9,25-pentatricontadiene and 9,27-heptatricontadiene (numbers indicate the position of double bonds in number of carbons from the carboxyl end of the hydrocarbon chain), have been discovered that vary between the Baja and Mainland populations. These components have been seen to influence the mating

success of males and to be sexually dimorphic, with sex specific differences depending on the population of origin (Etges & Ahrens 2001; Etges & Tripodi 2008; Markow & Toolson 1990; Stennett & Etges 1997). Etges *et al*. (2007) isolated QTLs for courtship song and premating reproductive isolation between flies from the Baja and Mainland populations on agria and organ pipe hosts. They showed that males had a higher mating success with Mainland females when they produced songs with shorter long-interpulse intervals, burst durations, and interburst intervals. Male mating success was influenced by a single QTL located on chromosome 2 near two desaturase genes, *desat1* and *desat2*, and there were extensive GEIs for courtship song. In a further study, Etges *et al*. (2009) isolated a number of QTLs influencing mating success and CHC composition, again finding extensive GEIs. However the three previously identified alkadiene CHC components seen in previous studies were not associated with mating success in this study.

Etges *et al*. (2009) discovered that mating success between Baja and Mainland populations was associated with just one genetic marker. This marker was situated near two desaturase genes, *desat1* and *desat2*, which are thought to be involved in CHC production (Wicker-Thomas & Chertemps 2010) and might underlie reproductive isolation between closely related *Drosophila* populations, although this has been disputed (Coyne & Elwyn 2006; Dallerac *et al*. 2000; Grillet *et al*. 2012; Ritchie & Noor 2004; Takahashi *et al*. 2001). Desaturase genes play a key role in pheromone synthesis. Thus, although the genetic basis of pheromone production and mating success is likely to be complex, the desaturase genes are good candidates for speciation genes in *D. mojavensis*, and are examined in Chapter 2 of this thesis.

The aim of this thesis is to examine the molecular basis of incipient speciation and population divergence. This wass primarily carried-out by examining incipient species of *D. mojavensis* populations, but I also examined the potential for epigenetic imprinting to contribute to divergence in populations of a freshwater fish, *Girardinichthys multiradiatus*. In the second chapter I examine genetic variation between populations of *D. mojavensis* in two candidate genes, previously identified through QTL mapping to reproductive traits and male mating success (Etges *et al*. 2007; Etges *et al*. 2009). These were the *desat1* and *desat2* genes that have been implicated in reproductive isolation in other *Drosophila* species. In Chapter 3 I perform a phylogeographic analysis of the *D. mojavensis* populations in order to ascertain the order of population divergence, timing of these splits and the levels of gene flow and migration that occurred during the process, thus characterizing the demographic context of this incipient speciation. In Chapter 4 I performed whole-transcriptome sequencing to examine cactus specific gene expression plasticity, and gene expression associated with mating success. This study identified genes (and their functions) that are expressed according to cactus specific mating success, which are potential candidates for speciation genes. Lastly, Chapter 5 examined the role of epigenetic imprinting in population divergence potentially due to sexual conflict in a species of freshwater fish. Through these studies I uncover genetic, expression and epigenetic variation involved in the divergence of closely related populations.

REFERENCES

Adams MD, Celniker SE, Holt RA, *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185-2195.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805-1814.

Arbuthnott D (2009) The genetic architecture of insect courtship behavior and premating isolation. *Heredity*, **103**, 15-22.

Aubin-Horth N, Renn SCP (2009) Genomic reaction norms: using integrative biology to understand molecular mechanisms of phenotypic plasticity. *Molecular Ecology*, **18**, 3763-3780.

Baker RJ, Bradley RD (2006) Speciation in mammals and the genetic species concept. *Journal of Mammalogy*, **87**, 643-662.

Baldwin JM (1896) A new factor in evolution. *American Naturalist*, 441-451.

Baldwin JM (1902) *Development and Evolution* MacMillan & Co., New York.

Baythavong BS (2011) Linking the spatial scale of environmental variation and the evolution of phenotypic plasticity: Selection favors adaptive plasticity in fine-grained environments. *American Naturalist*, **178**, 75-87.

Berlocher SH, Feder JL (2002) Sympatric speciation in phytophagous insects: Moving beyond controversy? *Annual Review of Entomology*, **47**, 773-815.

Boake CRB, DeAngelis MP, Andreadis DK (1997) Is sexual selection and species recognition a continuum? Mating behavior of the stalk-eyed fly *Drosophila heteroneura*. *Proceedings of the National Academy of Sciences, USA*, **94**, 12442-12445.

Boughman JW (2002) How sensory drive can promote speciation. *Trends in Ecology & Evolution*, **17**, 571-577.

Boughman JW, Rundle HD, Schluter D (2005) Parallel evolution of sexual isolation in sticklebacks. *Evolution*, **59**, 361-373.

Bradshaw A (1965) Evolutionary significance of phenotypic plasticity in plants. *Advances in Genetics*, **13**, 115-155.

Brazner JC, Etges WJ (1993) Pre-mating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. II. Effects of larval substrates on time to copulation, mate choice and mating propensity. *Evolutionary Ecology*, **7**, 605-624.

Butlin RK (2010) Population genomics and speciation. *Genetica*, **138**, 409-418.

Butlin RK, Debelle A, Kerth C, *et al*. (2012) What do we need to know about speciation? *Trends in Ecology & Evolution*, **27**, 27-39.

Carson HL, Kaneshiro KY, Val FC (1989) Natural hybridization between the sympatric hawaiian species *Drosophila silvestris* and *Drosophila heteroneura*. *Evolution*, **43**, 190-203.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research*, **70**, 155-174.

Cossins A, Fraser J, Hughes M, Gracey A (2006) Post-genomic approaches to understanding the mechanisms of environmentally induced phenotypic plasticity. *Journal of Experimental Biology*, **209**, 2328-2336.

Coyne JA, Elwyn S (2006) Does the *desaturase-2* locus in *Drosophila melanogaster* cause adaptation and sexual isolation? *Evolution*, **60**, 279-291.

Coyne JA, Orr HA (1989) Patterns of speciation in *Drosophila*. *Evolution*, **43**, 362-381.

Coyne JA, Orr HA (1998) The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **353**, 287-305.

Coyne JA, Orr HA (2004) *Speciation* Sinauer, Sunderland, Mass.

Coyne JA, Wicker-Thomas C, Jallon J-M (1999) A gene responsible for a cuticular hydrocarbon polymorphism in *Drosophila melanogaster*. *Genetics Research*, **73**, 189-203.

Crispo E (2007) The Baldwin effect and genetic assimilation: Revisiting two mechanisms of evolutionary change mediated by phenotypic plasticity. *Evolution*, **61**, 2469-2479.

Dallerac R, Labeur C, Jallon JM, *et al*. (2000) A Delta 9 desaturase gene with a different substrate specificity is responsible for the cuticular diene hydrocarbon polymorphism in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences, USA*, **97**, 9449-9454.

Darwin C (1859) *The origin of species*. John Murray, London.

De Queiroz K (2005) Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences, USA*, **102**, 6600-6607.

De Queiroz K (2007) Species concepts and species delimitation. *Systematic Biology*, **56**, 879-886.

De Queiroz K (2011) Branches in the lines of descent: Charles Darwin and the evolution of the species concept. *Biological Journal of the Linnean Society*, **103**, 19-35.

DeWitt TJ (1998) Costs and limits of phenotypic plasticity: Tests with predator-induced morphology and life history in a freshwater snail. *Journal of Evolutionary Biology*, **11**, 465-480.

DeWitt TJ, Scheiner SM (2004) *Phenotypic plasticity: functional and conceptual approaches* Oxford University Press, USA.

DeWitt TJ, Sih A, Wilson DS (1998) Costs and limits of phenotypic plasticity. *Trends in Ecology & Evolution*, **13**, 77-81.

Dobzhansky T (1936) Studies on hybrid sterilty. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics*, **21**, 113-135.

Dobzhansky T (1937) *Genetics and the Origin of Species* Columbia University Press, New York.

Dobzhansky T (1940) Speciation as a stage in evolutionary divergence. *The American Naturalist*, **74**, 312-321.

Dobzhansky T (1948) Genetics of natural populations. XVII. Experiments on chromosomes of *Drosophila pseudoobscura* from different geographic regions. *Genetica*, **33**, 588-602.

Dobzhansky T (1950) Mendelian populations and their evolution. *American Naturalist*, **84**, 401-418.

Dobzhansky T, Pavlovsky O (1966) Spontaneous origin of an incipient species in the *Drosophila paulistorum* complex. *Proceedings of the National Academy of Sciences, USA*, **55**, 727-733.

Dobzhansky T, Pavlovsky O (1967) Experiments of the incipient species of the *Drosophila paulistorum* complex. *Genetics*, **55**, 141-156.

Doebeli M, Dieckmann U, Metz JAJ, Tautz D (2005) What we have also learned: Adaptive speciation is theoretically plausible. *Evolution*, **59**, 691-695.

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203-218.

Etges WJ (1990) Direction of life history evolution in *Drosophila mojavensis*. In: *Ecological and Evolutionary Genetics of Drosophila*, pp. 37-56.

Etges WJ, Heed W (1987) Sensitivity to larval density in populations of *Drosophila mojavensis*: influences of host plant variation on components of fitness. *Oecologia*, **71**, 375-381.

Etges WJ (1992) Premating isolation is determined by larval substrates in cactophilic *Drosophila mojavensis*. *Evolution*, **46**, 1945-1950.

Etges WJ (1993) Genetics of host-cactus response and life-history evolution among ancestral and derived populations of cactophilic *Drosophila mojavensis*. *Evolution*, **47**, 750-767.

Etges WJ (1998) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. IV. Correlated responses in behavioral isolation to artificial selection on a life-history trait. *American Naturalist*, **152**, 129-144.

Etges WJ (2002) Divergence in mate choice systems: does evolution play by rules? *Genetica*, **116**, 151-166.

Etges WJ, Ahrens MA (2001) Premating isolation is determined by larval-rearing substrates in cactophilic *Drosophila mojavensis*. V. Deep geographic variation in epicuticular hydrocarbons among isolated populations. *American Naturalist*, **158**, 585-598.

Etges WJ, de Oliveira CC, Gragg E, *et al*. (2007) Genetics of incipient speciation in *Drosophila mojavensis*. I. Male courtship song, mating success, and genotype x environment interactions. *Evolution*, **61**, 1106-1119.

Etges WJ, De Oliveira CC, Noor MAF, Ritchie MG (2010) Genetics of incipient speciation in *Drosophila mojavensis*. III. Life history divergence in allopatry and reproductive isolation. *Evolution*, **64**, 3549-3569.

Etges WJ, de Oliveira CC, Ritchie MG, Noor MAF (2009) Genetics of incipient speciation in *Drosophila mojavensis*. II. Host plants and mating status influence cuticular hydrocarbon QTL Expression and G × E Interactions. *Evolution*, **63**, 1712-1730.

Etges WJ, Johnson WR, Duncan GA, Huckins G, Heed WB (1999) Ecological genetics of cactophilic *Drosophila*. In: *Ecology of Sonoran Desert plants and plant communities*. (ed. R R), pp. 164-214. University of Arizona Press, Tucson.

Etges WJ, Klassen CS (1989) Influences of atmospheric ethanol on adult *Drosophila mojavensis*: Altered metabolic rates and increases in fitness among populations. *Physiological Zoology*, **62**, 170-193.

Etges WJ, Over KF, De Oliveira CC, Ritchie MG (2006) Inheritance of courtship song variation among geographically isolated populations of *Drosophila mojavensis*. *Animal Behaviour*, **71**, 1205-1214.

Etges WJ, Tripodi AD (2008) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. VIII. Mating success mediated by epicuticular hydrocarbons within and between isolated populations. *Journal of Evolutionary Biology*, **21**, 1641-1652.

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729-1747.

Ferguson JWH (2002) On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society*, **75**, 509-516.

Ferveur JF (2005) Cuticular hydrocarbons: Their evolution and roles in *Drosophila* pheromonal communication. *Behavior Genetics*, **35**, 279-295.

Fitzpatrick BM (2012) Underappreciated consequences of phenotypic plasticity for ecological speciation. *International Journal of Ecology*, **doi:10.1155/2012/256017**.

Fitzpatrick BM, Fordyce JA, Gavrilets S (2009) Pattern, process and geographic modes of speciation. *Journal of Evolutionary Biology*, **22**, 2342-2347.

Fogleman JC, Danielson PB (2001) Chemical interactions in the cactus-microorganism-*Drosophila* model system of the Sonoran Desert. *American Zoologist*, **41**, 877-889.

Fordyce JA (2006) The evolutionary consequences of ecological interactions mediated through phenotypic plasticity. *Journal of Experimental Biology*, **209**, 2377-2383.

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, **10**, 241-251.

Gavrilets S (2003) Models of speciation: what have we learned in 40 years? *Evolution*, **57**, 2197-2215.

Gavrilets S (2005) "Adaptive speciation" - It is not that easy: A reply to Doebeli *et al*. *Evolution*, **59**, 696-699.

Gavrilets S, Hayashi TI (2005) Speciation and sexual conflict. *Evolutionary Ecology*, **19**, 167-198.

Ghalambor C, McKay J, Carroll S, Reznick D (2007) Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Functional Ecology*, **21**, 394-407.

Gillespie JH (1974) The role of environmental grain in the maintenance of genetic variation. *American Naturalist*, **108**, 145-151.

Gleason JM, Ritchie MG (1998) Evolution of courtship song and reproductive isolation in the *Drosophila willistoni* species complex: Do sexual signals diverge the most quickly? *Evolution*, **52**, 1493-1500.

Gleason JM, Ritchie MG (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.

Grant PR, Grant BR (1997) Genetics and the origin of bird species. *Proceedings of the National Academy of Sciences, USA*, **94**, 7768-7775.

Greenfield MD, Rodriguez RL (2004) Genotype-environment interaction and the reliability of mating signals. *Animal Behaviour*, **68**, 1461-1468.

Greenspan RJ, Ferveur J-F (2000) Courtship in *Drosophila*. *Annual Reviews of Genetics*, **2000**, 205-232.

Grether GF (2005) Environmental change, phenotypic plasticity, and genetic compensation. *American Naturalist*, **166**, E115-123.

Grillet M, Everaerts C, Houot B, *et al*. (2012) Incipient speciation in *Drosophila melanogaster* involves chemical signals. *Scientific Reports*, **2**, 224.

Hall JC (1994) The mating of a fly. *Science*, **264**, 1702-1714.

Harrison RG (1998) Linking evolutionary pattern and process: the relevance of species concepts for the study of speciation. In: *Endless Forms, Species and Speciation*. (eds. Howard DJ, Berlocher SH), pp. 19-31. Oxford University Press, Oxford.

Hausdorf B (2011) Progress toward a general species concept. *Evolution*, **65**, 923-931.

Heed W, Mangan R (1986) Community ecology of the sonoran desert *Drosophila*. In: *In: The Genetics and Biology of Drosophila*, pp. 311-345. Academic Press, London.

Hendry AP (2004) Selection against migrants contributes to the rapid evolution of ecologically dependent reproductive isolation. *Evolutionary Ecology Research*, **6**, 1219-1236.

Higashi M, Takimoto G, Yamamura N (1999) Sympatric speciation by sexual selection. *Nature*, **402**, 523-526.

Hollander J (2008) Testing the grain-size model for the evolution of phenotypic plasticity. *Evolution*, **62**, 1381-1389.

Hollocher H, Ting CT, Pollack F, Wu CI (1997a) Incipient speciation by sexual isolation in *Drosophila melanogaster*: Variation in mating preference and correlation between sexes. *Evolution*, **51**, 1175-1181.

Hollocher H, Ting CT, Wu ML, Wu CI (1997b) Incipient speciation by sexual isolation in *Drosophila melanogaster*: Extensive genetic divergence without reinforcement. *Genetics*, **147**, 1191-1201.

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA-Sequence Data. *Genetics*, **132**, 583-589.

Jennings JH, Etges WJ (2010) Species hybrids in the laboratory but not in nature: A reanalysis of premating isolation between *Drosophila arizonae* and *D. mojavensis*. *Evolution*, **64**, 587-598.

Johannes F, Porcher E, Teixeira FK, *et al*. (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genetics*, **5**, e1000530.

Johnson LJ, Tricker PJ (2010) Epigenomic plasticity within populations: Its evolutionary significance and potential. *Heredity*, **105**, 113-121.

Johnson WR (1980) *Chromosomal polymorphism in natural populations of the desert adapted species Drosophila mojavensis*. PhD thesis, University of Arizona.

Jones FC, Grabherr MG, Chan YF, *et al*. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55-61.

Kirkpatrick M, Ravigne V (2002) Speciation by natural and sexual selection: Models and experiments. *American Naturalist*, **159**, S22-S35.

Kokko H, Brooks R, McNamara JM, Houston AI (2002) The sexual selection continuum. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **269**, 1331-1340.

Lande R (1981) Models of speciation by sexual selection on polygenic traits. *Proceedings of the National Academy of Sciences, USA*, **78**, 3721-3725.

Lande R (2009) Adaptation to an extraordinary environment by evolution of phenotypic plasticity and genetic assimilation. *Journal of Evolutionary Biology*, **22**, 1435-1446.

Lande R, Kirkpatrick M (1988) Ecological speciation by sexual selection. *Journal of Thoretical Biology*, **133**, 85-98.

Le Rouzic A, Carlborg O (2008) Evolutionary potential of hidden genetic variation. *Trends in Ecology & Evolution*, **23**, 33-37.

Levine MT, Eckert ML, Begun DJ (2011) Whole-genome expression plasticity across tropical and temperate *Drosophila melanogaster* populations from Eastern Australia. *Molecular Biology and Evolution*, **28**, 249-256.

Levins R (1968) *Evolution in Changing Environments* Princeton University Press, Princeton, New Jersey.

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

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981-994.

M'Gonigle LK, Mazzucco R, Otto SP, Dieckmann U (2012) Sexual selection enables long-term coexistence despite ecological equivalence. *Nature*, **484**, 506-509.

Machado CA, Kliman RM, Markert JA, Hey J (2002) Inferring the history of speciation from multilocus DNA sequence data: The case of *Drosophila pseudoobscura* and close relatives. *Molecular Biology and Evolution*, **19**, 472-488.

Mackay TFC, Heinsohn SL, Lyman RF, *et al*. (2005) Genetics and genomics of *Drosophila* mating behavior. *Proceedings of the National Academy of Sciences, USA*, **102**, 6622-6629.

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, **10**, 565-577.

Mallet J (1995) A species definition for the modern synthesis. *Trends in Ecology & Evolution*, **10**, 294-299.

Mallet J (2001) Species, concepts of. *Encyclopedia of biodiversity*, **5**, 427-440.

Mank JE (2009) Sexual selection and Darwin's mystery of mysteries. *Science*, **326**, 1639-1640.

Markow TA (1991) Sexual isolation among populations of *Drosophila mojavensis*. *Evolution*, **45**, 1525-1529.

Markow TA, Toolson EC (1990) Temperature effects on epicuticular hydrocarbons and sexual isolation in *Drosophila mojavensis*. In: *Ecological and*

evolutionary genetics of Drosophila. (eds. F. BJS, T. SW, J. MR), pp. 315-331. Plenum, New York.

Massie KR, Markow TA (2005) Sympatry, allopatry and sexual isolation between *Drosophila mojavensis* and *D. arizonae*. *Hereditas*, **142**, 51-55.

Matzkin LM (2004) Population genetics and geographic variation of alcohol dehydrogenase (*Adh*) paralogs and glucose-6-phosphate dehydrogenase (*G6pd*) in *Drosophila mojavensis*. *Molecular Biology and Evolution*, **21**, 276-285.

Matzkin LM, Eanes WF (2003) Sequence variation of alcohol dehydrogenase (*Adh*) paralogs in cactophilic *Drosophila*. *Genetics*, **163**, 181-194.

Mayden RL (1997) A hierarchy of species concepts: the denouement in the saga of the species problem. In: *Species: the units of biodiversity*. (eds. Claridge MF, Dawah HA, Wilson MR), pp. 381-424. Chapman and Hall, New York.

Mayr E (1942) *Systematics and the origin of species*. Columbia University Press, New York.

McCarthy MI, Abecasis GR, Cardon LR, *et al*. (2008) Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, **9**, 356-369.

Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: Insights into functions. *Nature Reviews Genetics*, **10**, 155-159.

Morgan TH, Bridges CB (1916) Sex linked inheritance in *Drosophila*. *Carnegie Institution of Washington: Publication*, **237**, 1-88.

Muller HJ (1939) Reversibility in evolution considered from the standpoint of genetics. *Biology Reviews*, **14**, 185-268.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197-218.

Noor MAF (1997) How often does sympatry affect sexual isolation in *Drosophila*? *American Naturalist*, **149**, 1156-1163.

Noor MAF (2003) Genes to make new species. *Nature*, **423**, 699-700.

Noor MAF, Feder JL (2006) Speciation genetics: evolving approaches. *Nature Reviews Genetics*, **7**, 851-861.

Noor MAF, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences, USA*, **98**, 12084-12088.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009a) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375-402.

Nosil P, Harmon LJ, Seehausen O (2009b) Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution*, **24**, 145-156.

Nosil P, Schluter D (2011) The genes underlying the process of speciation. *Trends in Ecology & Evolution*, **26**, 160-167.

Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **365**, 185-205.

Orr HA, Masly JP, Presgraves DC (2004) Speciation genes. *Current Opinion in Genetics & Development*, **14**, 675-679.

Parsons KJ, Sheets HD, Skulason S, Ferguson MM (2011) Phenotypic plasticity, heterochrony and ontogenetic repatterning during juvenile development of divergent Arctic charr (*Salvelinus alpinus*). *Journal of Evolutionary Biology*, **24**, 1640-1652.

Paterson HEH (1985) The recognition concept of species. In: *Species and Speciation*. (ed. Vrba S), pp. 21-29. Transvsaal Museum, Pretoria, S. A.

Paterson HEH (1993) *Evolution and the recognition concept of species* Johns Hopkins University Press, Baltimore.

Pfeiler E, Castrezana S, Reed L, Markow T (2009) Genetic, ecological and morphological differences among populations of the cactophilic *Drosophila mojavensis* from southwestern USA and northwestern Mexico, with descriptions of two new subspecies. *Journal of Natural History*, **43**, 923-938.

Pfennig DW, Wund MA, Snell-Rood EC, *et al*. (2010) Phenotypic plasticity's impacts on diversification and speciation. *Trends in Ecology & Evolution*, **25**, 459-467.

Pigliucci M (2001) *Phenotypic plasticity: beyond nature and nurture* Johns Hopkins University Press.

Pigliucci M, Murren CJ, Schlichting CD (2006) Phenotypic plasticity and evolution by genetic assimilation. *Journal of Experimental Biology*, **209**, 2362-2367.

Pigllucci M (1996) How organisms respond to environmental changes: From phenotypes to molecules (and vice versa). *Trends in Ecology & Evolution*, **11**, 168-173.

Price TD, Qvarnstrom A, Irwin DE (2003) The role of phenotypic plasticity in driving genetic evolution. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **270**, 1433-1440.

Reed LK, Nyboer M, Markow TA (2007) Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. *Molecular Ecology*, **16**, 1007-1022.

Richards EJ (2009) Quantitative epigenetics: DNA sequence variation need not apply. *Genes & Development*, **23**, 1601-1605.

Rieseberg LH, Blackman BK (2010) Speciation genes in plants. *Annals of Botany*, **106**, 439-455.

Ritchie MG (2007) Sexual selection and speciation. *Annual Review of Ecology, Evolution and Systematics*, **38**, 79-102.

Ritchie MG, Noor MAF (2004) Evolutionary genetics: Gene replacement and the genetics of speciation. *Heredity*, **93**, 1-2.

Ritchie MG, Phillips SDF (1998) The genetics of sexual isolation. In: *Endless forms: species and speciation*. (eds. Howard DA, Berlocher S), pp. 291-308. Oxford University Press, Oxford, UK.

Ross C, Markow T (2006) Microsatellite variation among diverging populations of *Drosophila mojavensis*. *Journal of Evolutionary Biology*, **19**, 1691-1700.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336-352.

Rundle HD, Schluter D (2004) Natural selection and ecological speciation in sticklebacks. In: *Adaptive speciation*. (eds. Dieckmann U, Doebeli M, Metz J, Tautz D), pp. 192-209. Cambridge University Press, International Institute for Applied Systems Analysis.

Sabeti PC, Varilly P, Fry B, *et al*. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913-U912.

Schlichting CD, Levin DA (1986) Phenotypic plasticity - an evolving plant character. *Biological Journal of the Linnean Society*, **29**, 37-47.

Schlichting CD, Pigliucci M (1993) Control of phenotypic plasticity via regulatory genes. *American Naturalist*, **142**, 366-370.

Schlichting CD, Pigliucci M (1995) Gene regulation, quantitative genetics and the evolution of reaction norms. *Evolutionary Ecology*, **9**, 154-168.

Schlichting CD, Smith H (2002) Phenotypic plasticity: Linking molecular mechanisms with evolutionary outcomes. *Evolutionary Ecology*, **16**, 189-211.

Schluter D (2000) *The ecology of adaptive radiation* Oxford University Press, USA.

Schluter D (2001) Ecology and the origin of species. *Trends in Ecology & Evolution*, **16**, 372-380.

Schluter D (2009) Evidence for ecological speciation and its alternative. *Science*, **323**, 737-741.

Servedio MR (2001) Beyond reinforcement: The evolution of premating isolation by direct selection on preferences and postmating, prezygotic incompatibilities *Evolution*, **55**, 1909-1920.

Servedio MR (2004) The evolution of premating isolation: Local adaptation and natural and sexual selection against hybrids. *Evolution*, **58**, 913-924.

Servedio MR, Noor MAF (2003) The role of reinforcement in speciation: Theory and data. *Annual Review of Ecology Evolution and Systematics*, **34**, 339-364.

Servedio MR, Van Doorn GS, Kopp M, Frame AM, Nosil P (2011) Magic traits in speciation: 'magic' but not rare? *Trends in Ecology & Evolution*, **26**, 389-397.

Shaw KS, Parsons YM (2002) Divergence of mate recognition behavior and its consequences for genetic architectures of speciation. *American Naturalist*, **159**, S61-S75.

Simpson, G (1943) Criteria for genera, species and subspecies in zoology and paleontology. *Annals New York Academy of Science*, **44**, 145-178.

Smadja C, Galindo J, Butlin R (2008) Hitching a lift on the road to speciation. *Molecular Ecology*, **17**, 4177-4180.

Smith-Gill SJ (1983) Developmental plasticity: developmental conversion versus phenotypic modulation. *American Zoologist*, **23**, 47-55.

Sobel JM, Chen GF, Watt LR, Schemske DW (2010) The biology of speciation. *Evolution*, **64**, 295-315.

Spieth HT (1974) Courtship behavior in *Drosophila*. *Annual Review of Entomology*, **19**, 385-405.

Springer SA, Crespi BJ, Swanson WJ (2011) Beyond the phenotypic gambit: Molecular behavioural ecology and the evolution of genetic architecture. *Molecular Ecology*, **20**, 2240-2257.

Stennett MD, Etges WJ (1997) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. III. Epicuticular hydrocarbon variation is determined by use of different host plants in *Drosophila mojavensis* and *Drosophila arizonae*. *Journal of Chemical Ecology*, **23**, 2803-2824.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: Finding the genes underlying ecologically important traits. *Heredity*, **100**, 158-170.

Sun HM, Merugu S, Gu X, *et al*. (2002) Identification of essential amino acid changes in paired domain evolution using a novel combination of evolutionary analysis and in vitro and in vivo studies. *Molecular Biology and Evolution*, **19**, 1490-1500.

Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. *Nature Reviews Genetics*, **3**, 137-144.

Takahashi A, Ting CT (2004) Genetic basis of sexual isolation in *Drosophila melanogaster*. *Genetica*, **120**, 273-284.

Takahashi A, Tsaur SC, Coyne JA, Wu C-I (2001) The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila*

melanogaster. *Proceedings of the National Academy of Sciences, USA*, **98**, 3920-3925.

Templeton AR (1981) Mechanisms of speciation - a population genetics approach. *Annual Review of Ecology, Evolution and Systematics*, **12**, 23-48.

Thibert-Plante X, Hendry AP (2011) The consequences of phenotypic plasticity for ecological speciation. *Journal of Evolutionary Biology*, **24**, 326-342.

Ting CT, Takahashi A, Wu C-I (2001) Incipient speciation by sexual isolation in *Drosophila:* Concurrent evolution at multiple loci. *Proceedings of the National Academy of Sciences, USA*, **98**, 6709-6713.

Tirosh I, Weinberger A, Bezalel D, Kaganovich M, Barkai N (2008) On the relation between promoter divergence and gene expression evolution. *Molecular Systems Biology*, **4**, 1-11.

Turelli M, Barton NH, Coyne JA (2001) Theory and speciation. *Trends in Ecology & Evolution*, **16**, 330-343.

Turner GF, Burrows MT (1995) A model of sympatric speciation by sexual selection. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **260**, 287-292.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in Anopheles gambiae. *PLoS Biology*, **3**, 1572-1578.

Van Doorn GS (2004) *Sexual selection and sympatric speciation*. Ph.D thesis, University of Gronigen.

Via S (2001) Sympatric speciation in animals: The ugly duckling grows up. *Trends in Ecology & Evolution*, **16**, 381-390.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334-4345.

Waddington CH (1942) Canalization of development and the inheritance of acquired characters. *Nature*, **150**, 563-565.

Waddington CH (1953) Genetic assimilation of an acquired character. *Evolution*, 118-126.

Wasserman M, Koepfer HR (1977) Character displacement for sexual isolation between *Drosophila mojavensis* and *Drosophila arizonensis*. *Evolution*, **31**, 812-823.

Weatherhead PJ, Robertson RJ (1979) Offspring quality and the progeny threshold: "The sexy son hypothesis". *American Naturalist*, **113**, 201-208.

West-Eberhard MJ (1989) Phenotypic plasticity and the origins of diversity. *Annual Review of Ecological Systematics*, **20**, 249-278.

West-Eberhard MJ (1983) Sexual selection, social competition and speciation. *Quarterly Review of Biology*, **58**, 155-183.

West-Eberhard MJ (2003) *Developmental plasticity and evolution*, Oxford University Press, Oxford.

West-Eberhard MJ (2005a) Developmental plasticity and the origin of species differences. *Proceedings of the National Academy of Sciences, USA*, **102**, 6543-6549.

West-Eberhard MJ (2005b) Phenotypic accommodation: Adaptive innovation due to developmental plasticity. *Journal of Experimental Zoology, Part B: Molecular and Developmental Evolution*, **304**, 610-618.

Wicker-Thomas C, Chertemps T (2010) Molecular biology and genetics of hydrocarbon production. In: *Insect Hydrocarbons: Biology, Biochemistry, and Chemical Ecology*. (eds. Blomquist GJ, Bagnéres AG), pp. 53-74. Cambridge University Press, Cambridge.

Wiens JJ (2007) Species delimitation: New approaches for discovering diversity. *Systematic Biology*, **56**, 875-878.

Wilson EO, Brown Jr WL (1953) The subspecies concept and its taxonomic application. *Systematic Biology*, **2**, 97-111.

Wolf JBW, Lindell J, Backström N (2010) Speciation genetics: Current status and evolving approaches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 1717-1733.

Wright S (1931) Evolution in mendelian populations. *Genetics*, **16**, 97-159.

Wright S (1949) The genetical structure of populations. *Annals of Human Genetics*, **15**, 323-354.

Wu CI, Hollocher H, Begun DJ, *et al*. (1995) Sexual isolation in *Drosophila melanogaster*: A possible case of incipient speciation. *Proceedings of the National Academy of Sciences, USA*, **92**, 2519-2523.

Wu C-I, Ting C-T (2004) Genes and speciation. *Nature Reviews Genetics*, **5**, 114-122.

Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851-865.

Wund MA, Baker JA, Clancy B, Golub JL, Foster SA (2008) A test of the "Flexible stem" model of evolution: Ancestral plasticity, genetic accommodation, and morphological divergence in the threespine stickleback radiation. *American Naturalist*, **172**, 449-462.

Yukilevich R, True JR (2008) Incipient sexual isolation among cosmopolitan *Drosophila melanogaster* populations. *Evolution*, **62**, 2112-2121.

Zahavi A (1975) Mate selection - A selection for a handicap. *Journal of Theoretical Biology*, 205-214.

Zouros E, d'Entremont CJ (1974) Sexual isolation among populations of *Drosophila mojavensis* race B. *Drosophila Information Service*, **51**, 112.

Zouros E, d'Entremont CJ (1980) Sexual isolation among populations of *Drosophila mojavensis*: Response to pressure from a related species. *Evolution*, **34**, 421-430.

CHAPTER 2:

SEQUENCING OF TWO CANDIDATE GENES THAT POTENTIALLY UNDERLIE PHEROMONE VARIATION AND MATING SUCCESS BETWEEN DIVERGENT *DROSOPHILA MOJAVENSIS* POPULATIONS

ABSTRACT

Key to understanding the genetic basis of ecological speciation is discovering the genetic architecture of traits involved in adaptation and reproductive isolation. The search for speciation genes has been on-going for a number of years, but is fraught with difficulties involved with the analysis of complex traits that commonly are controlled by multiple loci. The *desaturase-1* (*desat1*) and *desaturase-2* (*desat2*) genes have previously been implicated in behavioural isolation between populations of *Drosophila melanogaster*. These desaturase genes are involved in the creation of double bonds during courtship pheromone synthesis and are good candidate genes for important pheromonal differences. This chapter examines these two candidate genes, which have also been previously implicated through QTL mapping to influence mating success in the cactophilic *Drosophila mojavensis*. *Desat1* and *desat2* were sequenced in the Baja and Mainland populations of *D. mojavensis* to examine population specific genetic variation. Sequencing encompassed the majority of introns and exons as well as a portion of the *desat2* promoter. Very little genetic variation was seen for both gene regions and no population-specific single nucleotide polymorphisms (SNPs) were fixed between the two populations. However, additional sequencing of the Santa Catalina Island population uncovered one nonsynonymous SNP between this population and the other two. Overall, results suggest that *desat1* and *desat2* are under purifying selection in both populations. A survey of methylation patterns in the *desat2* promoter was also carried out. DNA methylation is not thought to occur in *Drosophila* species, however low levels of DNA methylation were discovered at several different types of nucleotide motifs. These results highlight the difficulties of identifying key genetic changes underlying ecological isolation.

INTRODUCTION

An important step towards an integrated understanding of adaptation and speciation is to elucidate the molecular architecture of adaptive traits and reproductive isolation (Feder & Mitchell-Olds 2003). There are several complementary methods by which speciation genes and the mechanisms of their expression can be identified. Broadly, these can be split into candidate gene methods versus whole genome/transcriptome/proteome approaches, and include the examination of genomic variation (e.g. QTL studies, GWASs and genome scans), expressional and transcriptomic variation (microarrays, RNA-Seq) and proteomic variation (including protein-protein and DNA-protein interactions, e.g. ChIPseq). The candidate gene approach often follows a QTL or whole genome approach, yet narrowing down causative loci can be very challenging. However, the increasing utility and decreasing costs of new technologies such as high-throughput sequencing and microarrays now allow multiple lines of complementary evidence, increasing the confidence in the identification of a candidate gene. This is important because narrowing down a candidate locus and especially causative mutations is difficult, as is confirming the fitness effects of variation at such loci (Rockman 2012; Stinchcombe & Hoekstra 2008). Further, the molecular basis of adaptive evolution, which includes mRNA and protein expression, needs to be examined to gain a full understanding of the mechanisms by which phenotypes are expressed (Biron *et al*. 2006; Oleksiak *et al*. 2002). To begin to answer many of the key questions in adaptation and speciation we need to move towards a broader knowledge of the genetic basis underlying adaptive evolution, which includes identifying candidate genes and molecular mechanisms across a broad range of species and ecological settings.

Drosophila mojavensis is a cactophilic species with a well-studied ecology, that exists as four allopatric populations distributed across northwestern Mexico and the southwest United States, each living predominantly on a different cactus host (Fellows & Heed 1972; Ruiz *et al*. 1990) (Fig. 2.1). The Baja California population lives and breeds on pitaya agria cactus (*Stenocereus gummosus*), which is endemic to the Baja California peninsula whereas the Mainland population, distributed across the Sonoran desert and southern Arizona (Fig. 2.1), uses organ pipe cactus (*Stenocereus thurberi*) as its host. Studies have shown low but significant levels of behavioural isolation between these two populations, dependent on the rearing host cactus species (Brazner & Etges 1993; Etges 1992; Stennett & Etges 1997). The rearing of flies on different host cacti has been shown to influence multiple reproductive and life history traits, which demonstrate extensive levels of phenotypic plasticity and genotype-by-environment interactions across populations (Etges *et al*. 2007; Etges *et al*. 2010; Etges *et al*. 2009). Previous work to characterise putative candidate genes involved in ecological adaptation in *D. mojavensis* includes the alcohol dehydrogenase (*Adh*) genes, identified for their role in metabolism and detoxification linked to living on necrotic cactus tissue. The *Adh* genes are duplicate gene copies thought to play a role in adaptation to alcoholic environments (Mercot *et al*. 1994). In *D. mojavensis* these genes are expressed according to developmental stage, with *Adh-1* expressed in embryos and larvae, and *Adh-2* expressed in adults (Batterham *et al*. 1983). Signals of natural selection have been demonstrated in the coding sequence of *Adh-1*, between *D. mojavensis* and sister species *D. arizonae* (Matzkin & Eanes 2003) as well as between the Baja and Mainland populations of *D. mojavensis* (Matzkin 2004). Although the exact causative mutation, and thus the functional consequence of the amino acid substitution to protein function could not be ascertained, this is an example of

successful characterisation of a candidate gene underlying adaptation. However, a gene underlying the reproductive isolation between *D. mojavensis* populations has yet to be characterised.

Several recent studies have implicated candidate genes that function in the production of courtship pheromones to be involved in adaptation and speciation in *D. mojavensis* (Etges *et al*. 2007; Etges *et al*. 2010; Etges *et al*. 2009). Epicuticular hydrocarbons (CHCs) are long chain fatty acids that, in *Drosophila* species, act as non-volatile contact pheromones to elicit excitement in a potential sex partner (Ferveur *et al*. 1997; Ritchie & Noor 2004). CHC profiles are made up of blends of fatty acid components, which vary in composition across *Drosophila* species. Gas chromatography is used to examine CHC profiles and individual components are identified as peaks at an equivalent chain length of a previously identified molecular structure. For example, a chain length of 34.59 is known to be 8,24-tricontadiene, with the numbers indicating the position of the double bonds, counted from the carboxyl carbon. Adult *D. mojavensis* CHCs vary in chain lengths ranging from C_{29} to C_{39} as isomers such as alkanes, 2-methylalkanes, alkenes, methyl-branched alkenes and alkadienes, with the largest fraction of adult CHCs composed of C35 alkadienes (Etges & Jackson 2001; Toolson *et al*. 1990). The pattern of CHC expression across populations of *D. mojavensis* is complicated, showing sexual dimorphism and substrate specific differences (Stennett & Etges 1997). It is thought that *D. mojavensis* ultilise a suite of contact pheromones compared to the relatively simple differences seen in *D. melanogaster*, with 11 out of 20 CHC components being sexually dimorphic within *D. mojavensis* (Stennett & Etges 1997). The three principle alkadienes: 8,24-tricontadiene (chain length $C_{32.63}$), 9,25-pentatricontadiene ($C_{34.59}$) and 9,27- heptatricontadiene ($C_{36.5}$) have been specifically implicated in mate

recognition between the Mainland and Baja populations. These three alkadienes have been shown to influence mating success, suggesting a role for CHCs in the behavioural isolation between the Baja and Mainland populations (Etges & Ahrens 2001).

Recent quantitative trait loci (QTL) analyses identified a potential role for two desaturase genes, *desat1* and *desat2*, in the evolution of sexual isolation and life history traits between the Baja and Mainland groups (Etges *et al*. 2007; Etges *et al*. 2010; Etges *et al*. 2009). These studies examined CHC, courtship song and life history trait variation in Baja and mainland crosses, along with courtship trials to determine the mating success of F_2 males with mainland females. Although several regions of the genome influenced these phenotypes, only one marker (Dmoj2_1603a) was associated with mainland specific variation in CHCs and courtship song. This marker was also associated with male mating success, where mainland alleles significantly increased success with mainland females. The location of this marker was near two Δ9 desaturase genes: *desat1* and *desat2*. *DESAT1* is involved in the first desaturation of fatty acids using palmitate as its substrate, creating double bonds at the ω7 position of hydrocarbon chains (Gleason *et al*. 2005). *DESAT2* uses myristate as its substrate creating double bonds at the ω5 position (Dallerac *et al*. 2000; Fang *et al*. 2002). *Desat2* has been previously linked to sexual isolation in *D. melanogaster* where a 16bp deletion in the promoter region correlated with reproductive isolation between incipient groups (Dallerac *et al*. 2000; Takahashi *et al*. 2001). *Desat1* has been shown to play a role in pheromone production and perception in *Drosophila* (Bousquet *et al*. 2012; Bousquet & Ferveur 2012). However, the role of desaturase genes in the adaptation and sexual isolation of *D. melanogaster* populations has been disputed. Work replicating previous experiments failed to find allelic variation in

desat2 linked to climatic adaptation, and found inconsistent observations between wild-type populations and transgenic lines when examining sexual isolation (Coyne & Elwyn 2006). Further, the expression of *desat1* and *desat2* was not correlated to mating patterns between African and cosmopolitan strains that demonstrated sexual isolation and pheromonal variation (Grillet *et al*. 2012). Thus, the role of *desat1* and *desat2* is *Drosophila* sexual isolation is far from clear. However, their role in pheromone production still makes them viable candidates for speciation genes underlying the incipient speciation of the Baja and Mainland *D. mojavensis* populations.

In order to identify population-specific genetic variation in the desaturase loci, *desat1* and *desat2* were sequenced across the Baja and Mainland populations. The main aims of this study were to 1) Look for signatures of natural selection in the coding regions, 2) Identify any fixed population-specific mutations that might explain CHC variation 3) Identify the locations of such mutations (coding or non-coding) to infer the targets of selection in these genes. Through these tests the potential targets of selection can be identified allowing inference of the evolutionary processes that govern their evolution in *D. mojavensis* populations. An additional aim of this study was to assess if DNA methylation occurs in candidate speciation genes in *D. mojavensis*. The promoter region of *desat2* has been previously implicated in influencing pheromone expression and sexual isolation in *Drosophila* (Dallerac *et al*. 2000; Takahashi *et al*. 2001), thus the promoter region of *desat2*, rather than *desat1*, was examined for DNA methylation. The presence of DNA methylation in *Drosophila* has been a controversial subject, yet studies have shown that DNA methylation occurs in the *D. melanogaster* genome at very low levels (Lyko *et al*. 2000). However, there is little evidence of its functionality. First, I confirmed the

presence of DNA methylation in the *desat2* promoter region. Second, due to the evidence that epigenetic marks can be environmentally induced (Richards 2006; Richards 2008), I looked for environmental variation in methylation across flies raised on organ pipe cactus and normal lab food. Lab food was used because it is known to have very strong effects on pheromone production compared to cactus substrates (Stennett & Etges 1997). Given the potential role of regulatory features in *desat2* expression in *D. melanogaster*, any consistent pattern of DNA methylation in promoter regions of this candidate gene may indicate its function in gene expression regulation.

Fig. 2.1 Map of the sample site locations used in this study. Sample sites are black dots with full name and abbreviation. Dashed lines represent population boundaries with population name in italics. Solid black lines denote the approximate boundaries of the Mojave and Sonoran deserts. In total 38 sequences were obtained from the Baja population and 103 from the Mainland population.

METHODS

Sampling and DNA extraction

Individuals were sampled from laboratory strains derived from field populations collected in 2008 and 2009. The sampling scheme encompassed two sub-populations from the Baja peninsula (Punta Prieta and San Quentin), and four from the Mainland

population (Punta Onah, Organ Pipe National Monument, Las Bocas and Puerto Choyuda; Fig. 2.1). Each sample was sequenced for *desat1* and *desat2* gene regions (see Fig. 2.2), which included introns, exons and ~500bp of the upstream promoter region of *desat2*. Additionally, several Santa Catalina Island individuals were sequenced for the *desat2* promoter as initial comparisons demonstrated interesting patterns of genetic variation in this region.

DNA was extracted from flies using a basic salt extraction protocol. Lysis 'Squishing' buffer was prepared with 0.5ml of 1M TrisHCL pH 8.2, 0.2ml of 0.5M EDTA, 0.25ml of 5M NaCl and 49.05ml water, for a 50ml total. A volume of 10μL of proteinase K [20mg/ml] was added to 990μL of squishing buffer. Each fly was put into a 1.5ml eppendorph tube and placed in the freezer for 5 minutes. 50μL of buffer was then added to the tube and the tissue was homogenized. Tubes were then incubated at 43°C overnight, boiled for two minutes to denature the proteinase K, and stored at -20°C.

DNA sequencing

Extracted DNA was amplified by PCR using primers designed in Primer3 (Rozen & Skaletsky 2000). Three overlapping primer sets were designed for each gene, to ensure the whole gene region was covered (Table 2.1). Sequencing was carried out using the forward primer for each set except *desat2* primer set 3, which was also sequenced with the reverse primer in order to obtain the full genomic sequence. The last *desat1* primer set was not sequenced for the reverse primer and thus some of exon 4 was not obtained. In total this produced four alignments encompassing *desat2* and three covering the majority of *desat1* (Fig. 2.2). PCR was carried out with the following protocol, per reaction; 5 μL of 10× ammonia buffer, 1 μL of 10mM dNTPs,

1 μL of 50mM $MgCl_2$, 1.2 μL of each primer at 30 pM/μL, 0.3 μL of 5U *Taq*, 9.9 μL of Q-Solution (Bioline), 29.5 μL of water and 1 μL of DNA for a 50 μL reaction volume. Thermoprofiles were; 94°C for 3 minutes followed by 30 cycles of 94°C for 30 seconds, primer specific T_A °C (Table 2.1) and 72°C for 30 seconds. Final extension was at 72°C for 5 minutes. PCR products were then purified using the MSB Spin PCRapace clean up kit (Thistle Scientific). PCR fragments were sequenced on an ABI 3730 sequencer after a BigDye reaction.

Table 2.1. Primer details for all seven alignments produced (see Fig. 2.2 for positions). All primer sets were sequenced using the forward primer, and *desat2* primer set 3 was also sequenced with the reverse primer in order to obtain the last exon.

Primer set	Primer sequence (5' to 3')	T_A (°C)
DS1PS1	F: TGAAGCGTGTCAAGTTCAGC R: GCATTGTTTGGGCATGCTA	59
DS1PS2	F: CATATCATATAGTCCAATAGATCAGACACG R: GCCAGAGACAACATAAAGAGCG	57
DS1PS3	F: GAGTATATGCGAAACAAAGCTGAAGATG R: ATTCTGGCGACCGTGCGAG	51
DS2PS1	F: TTTGCTCGGTTTAACTTGCAT R: ATTGTTTGCCATTTGGCTTG	60
DS2PS2	F: ATACCGCATACCAAATTTCAAAGATCTACA R: CCTTGTCTAGACTTGCATAATTACTGAGA	55
DS2PS3	F: CTCCATCCGTCCTTGATTTCATCAACT R: CTACTCTAGAAAATACAACCCGCTAAGCAT	55

DS, desaturase gene; PS, primer set; T_A (°C), annealing temperature in degrees centigrade.

Bisulfite sequencing of the desat2 *promoter*

A common method for examining methylation patterns in DNA sequences is to chemically treat genomic DNA, converting cytosines (C) to thymines (T), followed by cloning and sequencing (Jiang *et al*. 2010). This initial step is useful for verifying the presence of DNA methylation. This protocol was first carried out on one individual to establish the presence of patterns of DNA methylation, and was

followed by direct sequencing of PCR products for several individuals from the Mainland population.

Bisulfite sequencing involves the deamination of C to T using a buffered solution of sodium bisulfite. Methylated Cs are protected from this reaction by the methyl group and thus appear in the sequence as a C. It has been demonstrated that some species do not adhere strictly to the CpG rule of DNA methylation but also demonstrate methylation at CpA and CpT positions (Kunert *et al*. 2003). Therefore primers for methylation sequencing were designed using Methprimer and based on CpG islands, however these primers were picked to also target potential CpA islands (see Supplementary Material for details). The *desat2* promoter was examined due to its putative functional role in pheromone expression in *D. melanogaster* (Dallerac *et al*. 2000; Takahashi *et al*. 2001). Primers were designed in the promoter region of *desat2*, amplifying 445bp fragments spanning a region of 600bp-450bp upstream of the translation initiation site (TIS), coincident with CpA and CpG islands discovered in this region by performing CpA and CpG island searches (Fig. S2.2). CpA/CpG island discovery was performed using a modified Perl program script from the online software CpGcluster (http://bioinfo2.ugr.es/CpGcluster/), using default settings. One individual from the Mainland population (PC sample site; Fig. 2.1) was used for this initial survey.

DNA was extracted using a Qiagen tissue lysis buffer and treated with a buffered solution of sodium bisulfite with a Qiagen EpiTect Plus DNA Bisulfite kit, and incubated for five hours. PCR was carried out using the bisulfite sequencing primers described above, with hot-start *Taq* and Qiagen Q-solution for increased specificity of primer annealing. The PCR mix was; 20μL of Qiagen PCR mix (an optimized solution of dNTPs, ammonia buffer and $MgCl_2$), 0.4μL each of forward

and reverse primer, 0.6μL of DNA, 15.6μL of water and 5μL of Q-solution for a total of 42μL. The thermocycling profile consisted of an activation step of 15 minutes at 95°C, followed by 40 cycles of 94°C for 30 seconds, 57.5°C for 90 seconds and 72°C for 90 seconds, with final elongation at 70°C for 6 minutes. The PCR product was then checked on a 2% agarose gel to ensure a single fragment had been amplified before being purified using a MSB Spin PCRapace kit (Thistle Scientific). A TOPO TA cloning kit (Invitrogen) was used to clone the PCR product and nine colonies were picked for purification and sequencing. Bisulfite sequencing was also carried out directly from the PCR product (i.e. without cloning) on three male and three female flies from the Mainland (OPNM) population, raised on both organ pipe cactus and laboratory food (see Methods in Chapter 4) in order to examine potential methylation variation.

Sequence editing and Analysis

Raw chromats were edited by hand in Geneious (Drummond *et al*. 2011) and aligned using MAFFT (Katoh *et al*. 2002). Indels were encountered in some of the sequences covering non-coding regions, which caused the entire sequence to appear to be heterozygous at most positions. Sequence traces showed little background noise and thus heterozygous positions were called in Geneious using the IUPAC ambiguity code. These superimposed sequences were then separated using Indelligent (Dmitriev & Rakitov 2008), which splits the two sequences and identifies the indel position. For these indel-split sequences, one was chosen at random for analysis. Sequences with heterozygous SNPs were phased using PHASE v2.1.1 (Stephens & Scheet 2005; Stephens *et al*. 2001) with input files prepared using SeqPHASE (Flot 2010). Once phased, one of the two reconstructed haplotypes was chosen at random for inclusion

in downstream analyses. The assembly plugin in Geneious was used to assess the coverage across both genes. For this, sequences were treated as if they were long reads and assembled against the desaturase genes from the sequenced genome of *D. mojavensis*.

Genetic diversity and divergence tests were carried out in DNAsp v5.10.01 (Librado & Rozas 2009). Per site pairwise nucleotide diversity (π_S), and Tajimas's D (1989) were calculated for each of the seven alignments (Nei 1987). DNA divergence by population in DNAsp was used to identify single nucleotide polymorphisms (SNPs), their position (coding or non-coding) and to examine population specific genetic variation. TFSEARCH v1.3 (Akiyama 1995) was used for transcription factor binding site searches associated with certain SNPs and one indel sequence (see Results). TFSEARCH uses the Transfac database (Heinemeyer *et al*. 1998) to search for transcription factor binding site matches to a given input sequence and gives each match a score. Each transcription factor binding site search was carried out using the arthropod specific database and 10bp were taken up and downstream of the focal SNP for examination, using a threshold score of 85%. Due to a population specific SNP in the translation initiation site of *desat2* (see Results), TIS Miner (Liu & Wong 2003) was used to confirm the position of the TIS in the *desat2* gene. TIS miner identifies potential translation initiation sites and compares them to a training data set, producing a similarity score to the training set consensus sequence.

Bisulfite sequenced clones were trimmed to remove vector sequence, aligned in Geneious and checked for accuracy by eye. The number of different methylated motifs were quantified to 1) determine if methylation occurs in the *desat2* gene promoter and 2) which motifs are methylated.

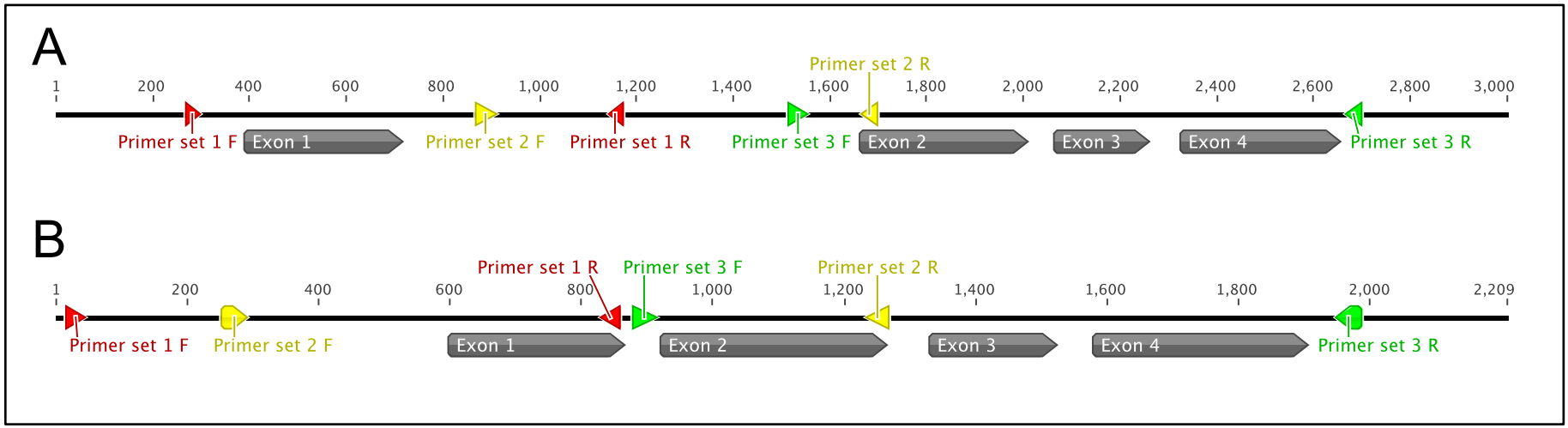


Fig. 2.2 Primer positions for sequencing of (a) *desat1* and (b) *desat2*. Exons are indicated in grey bars along with positions and directions of each primer pair from Table 2.1.

RESULTS

Genetic variation in the desaturase genes

In total 146 sequences (including five Santa Catalina Island samples) were obtained with an average length of 479bp (Table 2.2). Sequences covered the majority of gene regions for both *desat1* and *desat2*, with only exon 4 of *desat1* not surveyed (Fig. S2.1). At least five individuals were sampled from the Baja population per alignment, and always more than 10 for the Mainland population. Although sample sizes were fairly low (in order to survey across the genes), sampling several sub-populations should provide enough evidence of population-specific variation.

Table 2.2. Summary of sequence alignments across *desat1* and *desat2* for Baja and Mainland populations.

Gene region	Length	*N*	Total polymorphic sites	Informative sites	π_S	*D*
DS1PS1	632	22	5	1	0.00145	-1.3869
DS1PS2	503	22	11	4	0.00362	-1.3803
DS1PS3	582	19	4	2	0.00192	-1.1691
DS2PS1	339	21	32	11	0.01699	-1.3796
DS2PS2	408	15	8	5	0.00496	-0.6891
DS2PS3_F	392	24	8	6	0.00413	-0.7828
DS2PS3_R	500	18	9	5	0.00563	0.2714
Total	3356	141	77	34		
Mean	479	20	11	5		

DS, desaturase gene followed by number; PS, primer set followed by number; _F indicates sequenced with the forward primer; _R, indicates sequenced with the reverse primer; *N*, sample size; Informative (segregating) sites are positions with a minimum of two nucleotides that appear at least twice; π_S, per site nucleotide diversity (Nei, 1987); *D*, Tajima's D statistic (1989).

Nucleotide diversity across the desaturase genes was generally low, as was the average number of informative positions (i.e. positions with at least two SNPs that are present at least twice) across all alignments (Table 2.2). The nucleotide diversity across the desaturase genes was compared to the same diversity measure across 25 independent loci from the *D. mojavensis* genome. These loci represented a genome-

wide sample covering coding and noncoding regions and all chromosomes. Although a number of other factors may influence genetic diversity (for example recombination rate variation, proximity to transposable elements and inversions polymorphism amongst others), this sampling gave an approximate genome-wide average diversity for comparison across populations. Nucleotide diversity in the desaturase genes (median = 0.00413) was significantly lower than the average of the 25 loci (median = 0.0104, see Chapter 3 and Machado *et al*. 2007 for all loci used; Wilcoxon rank sum test, *p*-value <0.01). Tajimas's D (1989) tests for the influence of population history and natural selection on nucleotide diversity in sequence data, and significant deviations are taken as evidence for a lack of neutrality. The majority of alignments showed negative values for this neutrality test, none of which were significant, with *desat1* showing particularly strong negative values (Table 2.2). This may indicate the presence of purifying selection, however, small numbers of sampled alleles and a strong population structure can often reduce the power of neutrality tests (Nielsen 2001; Simonsen *et al*. 1995). The majority of polymorphic positions discovered were singleton SNPs, with none of the variable positions demonstrating fixed differences between populations. Several positions were fixed in one population and not the other, however these sites were all located in non-coding or synonymous genetic regions (i.e. non-coding or third codon coding sites). Thus no tests for selection could be undertaken due to the complete lack of non-synonymous mutations. Although no two SNPs were fixed between populations, 11 positions across all alignments were fixed for one allele in one population and at a high frequency for the alternative allele (defined as >50% of occurrences of the alternative SNP in the polymorphic population). These positions might potentially represent on-going fixation of alternative SNPs between populations.

These segregating sites were predominantly located in intron and third codon exon positions, indicating that any on-going fixation in one population is likely due to genetic drift and represents the independent evolutionary history of these populations. However, the *desat2* promoter region might be under divergent selection if SNPs lie in transcription-factor binding sites. To test this, five polymorphic positions discovered in the *desat2* promoter, with one allele fixed in one of the two populations and >50% alternative SNPs in the other, were examined for known transcription factor binding sites in arthropods using TFSEARCH. This search demonstrated no correspondence of any of the five positions to known transcription factor binding sites. Given their position in non-coding DNA, along with no detectable functional significance, it seems likely that these sites are evolving through genetic drift. No fixed differences were discovered between the Baja and Mainland population, however three fixed differences were seen between the Santa Catalina Island population and other two populations.

Table 2.3. Frequencies of polymorphic SNP positions between the Baja and Mainland populations.

Type of variation	DS1PS1	DS1PS2	DS1PS3	DS2PS1	DS2PS2	DS2PS3_F	DS2PS3_R
Poly. in Baja, mono. in Mainland	1	3	1	9	4	2	2
Poly. in Mainland, mono. in Baja	3	8	1	18	2	6	5
Fixed between populations	0	0	0	0	0	0	0
Shared between populations	1	0	2	5	2	0	2

Poly., polymorphic; mono., monomorphic; DS, desaturase gene; PS, primer set; _F indicates sequenced with the forward primer; _R, indicates sequenced with the reverse primer.

Fixed population differences were discovered between the Santa Catalina population and the other populations in the *desat2* promoter region, a first codon position in exon

1 and a third codon position in exon 1. The first codon position SNP caused a nonsynonymous change in the first amino acid of the protein in the Mainland and Baja populations, changing the methionine to a leucine (Fig. 2.3). Because methionine is always the first amino acid in any polypeptide, this mutation might mean either a population specific translation initiation site for *DESAT2*, or that the official genome release is incorrectly annotated and the true TIS is at another position. The latter explanation would lead to a nonsynonymous mutation in the first exon of *desat2* between the Santa Catalina Island and Baja/Mainland sequences. Thus a TIS search was carried out to identify the most likely TIS in both the Santa Catalina Island consensus sequence, and the Baja/Mainland consensus sequence. This search was carried out on each sequence from 137bp upstream of the previously annotated TIS. TIS miner identified four potential TISs in the Santa Catalina Island consensus sequence, two of which had in-frame stop codons in the 100bp downstream of the TIS, suggesting that these TISs did not produce functional proteins. The remaining two positions included the officially annotated TIS and a new TIS located 36bp upstream of the original. This new TIS had a higher score to the consensus TIS than the official annotation (0.179 compared to 0.002 respectively; Table S2.1), suggesting it was a more likely TIS. A search of the Baja/Mainland consensus sequence confirmed this new upstream TIS, thus for all populations the most likely true TIS lies 36bp upstream of the originally annotated protein (Table S2.1). The ATG > CTG mutation is therefore a fixed, nonsynonymous mutation between the Santa Catalina Island and other two populations. This mutation changes the methionine in the Santa Catalina Island population into a leucine in the Baja and Mainland groups (Fig. 2.3).

Although some sequences demonstrated indels, the majority of these were single or dinucleotide indels that occurred at low frequency, and showed no

population patterns in presence or absence. However, one indel located in the promoter region of *desat2* was present in all Baja and Mainland samples and absent in all Santa Catalina Island individuals (Fig. 2.4). This indel was 35bp long and a transcription factor binding site search discovered a binding site for chorion factor 2 (*Cf2*), which is always absent in the Santa Catalina population. *Cf2* functions in the regulation of muscle-related gene expression (Garcia-Zaragoza, et al, 2008), specifically in the control of indirect flight muscle development (Gajewski and Schultz, 2010).

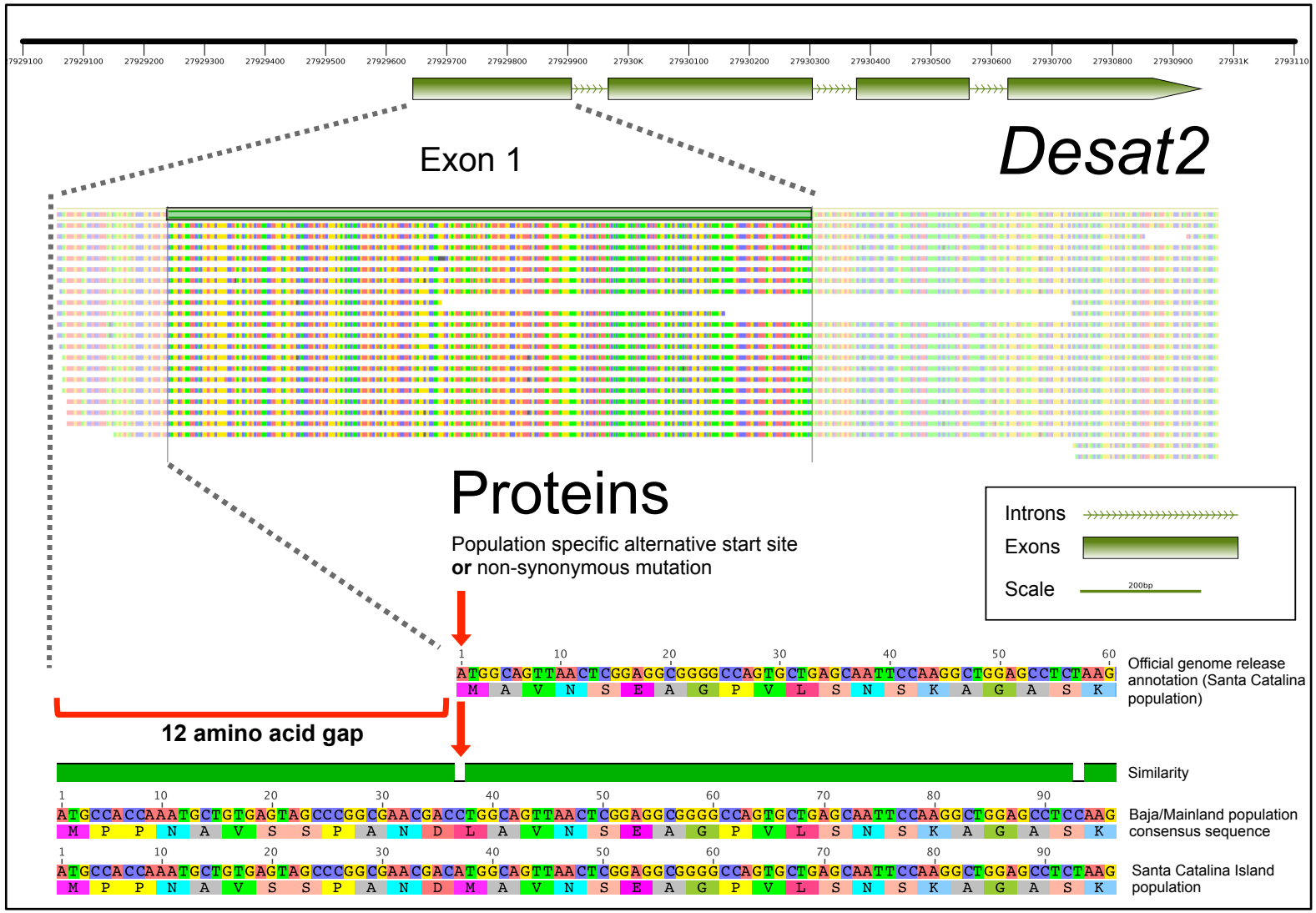


Fig. 2.3 Exon 1 coding sequence of *desat2*, showing the position of the officially annotated translation initiation site (TIS) and the TIS discovered with TIS miner, 36bp upstream. Identification of this new TIS uncovered a population specific SNP causing a change in the protein sequence of *DESAT2*.

Fig. 2.4 *Desat2* assembly showing coverage of the gene region, and population specific indel with predicted chorion factor 2 (*CF2-II*) binding site and absence of this site in the Santa Catalina Island population.

DNA methylation variation in desat2

DNA methylation was discovered in the promoter region of *desat2* at very low levels, with little indication of functional patterns. Because the whole body of one individual was surveyed, the cloned sequences represent the population of different cell types within this individual. Functional methylation is expected to appear as positions that are consistently methylated across all sequences. However, although there were clusters of methylated positions across sequences, no single nucleotide site was consistently methylated, indicating a random pattern of DNA methylation. The most frequently methylated motif was CpT followed by CpA and CpG, and CpC positions demonstrated no methylation (Table S2.2). Direct sequencing of PCR products was unsuccessful, except for three females raised on organ pipe cactus. These sequence traces had minor levels of background noise but were clean enough to identify the presence of methylation. Direct sequences demonstrated methylation at all nucleotide positions discovered in the sequenced clones, as well lower levels of methylation at other positions, most noticeably at CpA motifs.

DISCUSSION

Sequencing of two desaturase genes was carried out to examine population specific patterns of genetic variation across the introns and exons of *desat1* and *desat2*, and the promoter region of *desat2*, for the Baja California and Mainland populations of *D. mojavensis*. *Desat1* and *desat2* have been implicated in mating success and CHC expression divergence between the Baja and Mainland populations of *D. mojavensis* (Etges *et al*. 2009). However, little genetic variation was detected in both gene regions. Although a number of polymorphic positions were discovered, these were all located in intron and third codon position sites and thus are unlikely to have functional

consequences. None of the mutations discovered caused a nonsynonymous change in either amino acid sequence of the desaturase proteins, and thus no tests for patterns of selection could be undertaken. The lack of genetic variation in general, and complete lack of nonsynonymous mutations suggests that divergent natural selection is not acting on these genes between the Baja California and Mainland populations.

Previous QTL analyses and studies from other species suggested the desaturase genes were potential candidate loci, however the results presented here demonstrate that a direct role is perhaps unlikely. The QTL analyses of Etges *et al* (2007; 2009 & 2010) found that marker Dmoj2_1603a was consistently associated with CHC, courtship song and mating success variation. This marker was located near the *desat1* and *desat2* genes, and thus these genes were identified as likely candidate speciation genes. However, due to the high level of recombination in *D. mojavensis* it was not possible for these three studies to precisely identify genomic regions associated with phenotypic variation, and candidate genes were instead identified through their genetic association with particular genomic markers. Therefore, although the desaturase genes were good potential candidates, other regions in genetic linkage with marker Dmoj2_1603a might underlie the sexual isolation between *D. mojavensis* populations. These regions could include other *cis* or *trans* regulatory elements, such as enhancers, that might act to regulate the expression of *desat1* or *desat2*. Alternatively, they might be entirely different gene encoding regions, located near the desaturase loci and marker Dmoj2_1603a.

A high level of phenotypic divergence between genetically similar species suggests the importance of selection on gene expression variation, and genetic variation in transcription factor binding sites is a contributor to expression variation (King & Wilson 1975; Oleksiak *et al*. 2002). Although higher levels of genetic

variation were discovered in the *desat2* promoter region, none of the polymorphic positions were associated with known functional motifs. Thus it seems unlikely that divergent selection has acted on regulatory regions in this gene. Recent studies have shown that the *D. melanogaster desat1* promoter region contains at least five regulatory elements, producing five transcripts, which regulate gene expression in neural and non-neural tissues to coordinate the production and detection of courtship pheromones (Bousquet & Ferveur 2012; Bousquet *et al*. 2012). The promoter region of *desat1* was not surveyed here and thus variation in regulatory elements influencing *desat1* expression in *D. mojavensis* might exist, however recent evidence disputes the role of both desaturase genes in the sexual isolation of *D. melanogaster* populations (Grillet *et al*. 2012). The majority of polymorphic sites across both genes were located in non-coding and synonymous positions. Further, the lack of genetic variation over 3Kb of both gene regions along with a negative Tajima's D across most of these genes suggests the influence of purifying selection. A significantly lower level of nucleotide diversity was seen across the desaturase genes when compared to multiple independent loci indicating the action of background selection, which depletes genetic variation linked to deleterious mutations under purifying selection. Although it should be noted that this test was not corrected for a number of additional influences on nucleotide diversity such as physical location in the genome, recombination rate variation and inversion polymorphisms. Further, a selective sweep would also account for this reduction in nucleotide variation across the genes and further tests are necessary to distinguish this from background selection. Regardless, it is likely that a balance of purifying selection, and genetic drift are the primary forces governing the evolution of the desaturase genes in the Baja and Mainland populations of *D. mojavensis*.

Several fixed differences were discovered between the Santa Catalina Island and other two populations, including one nonsynonymous mutation in exon 1 and a large indel in the promoter region. The Santa Catalina population lives on an island off the west of California and uses opuntia cactus, *Opuntia demissa*, as a host plant (Ruiz *et al*. 1990). The four populations of *D. mojavensis* differ little in their morphology, and external characteristics do not differentiate the Santa Catalina group from the other three, the only exception to this being the shape of the aedeagus, which is population specific (Pfeiler *et al*. 2009). The Santa Catalina Island population does not demonstrate any significant premating reproductive isolation from any other *D. mojavensis* population (Markow 1991) or postzygotic isolation, although the latter has not been thoroughly examined (Mettler 1963; Reed & Markow 2004; Ruiz *et al*. 1990). Population genetic and phylogenetic analyses demonstrate that Santa Catalina is a panmictic population with strong differentiation from the other populations, showing little signal of gene flow between them (Machado *et al*. 2007; Reed *et al*. 2007; Ross & Markow 2006). Therefore it is interesting that *desat2* demonstrates strong population-specific signals of potentially functional variation, however the influence of this variation on CHC expression remains unclear, and CHCs in this population have not been characterised. A prospective transcription factor binding site was discovered in a 35bp indel that functions in flight muscle development. The flight muscles are known to provide power for courtship song production in *Drosophila* (Ewing 1979), potentially linking *desat2* to courtship song production. Given the small size of Santa Catalina Island there is the potential for increased genetic drift. Thus it remains to be seen how significant these observed functional population differences are, and what influence, if any, they have on the resulting phenotypes.

Very little epigenetic variation was seen in the *desat2* promoter region, suggesting low levels of DNA methylation in *D. mojavensis*. There has been some controversy in the last two decades concerning the extent of DNA methylation that occurs in the *Drosophila* genome. Early work found no evidence for DNA methylation in *Drosophila* genes (Rae & Steele 1979), however genome-wide DNA methylation was later discovered at very low levels in adults and slightly higher levels in embryos (Lyko *et al*. 2000). There is some evidence of a functional DNA methylation system in *Drosophila* (Lyko *et al*. 1999) yet the two main DNA methylation genes (*Dnmt1* and *Dnmt3*) are absent in the *Drosophila* genome (Krauss & Reuter 2011), suggesting a limited presence and functional importance for DNA methylation. However, *Dnmt2* is present and has been shown to be involved in the methylation of tRNAs (Schaefer *et al*. 2010) suggesting that it does have a functional role in epigenetic modifications, yet RNA interference of *Dnmt2* expression had no effect on fly development even though it resulted in a loss of DNA methylation (Kunert *et al*. 2003). Here, DNA methylation was discovered at low levels in different frequencies for each dinucleotide motif. Surprisingly, CpG motifs were very lightly methylated compared to CpT, which was the most frequently methylated dinucleotide. Both CpA and CpT motifs were methylated and studies have linked CpA/CpT methylation to the action of *Dnmt2* (Kunert *et al*. 2003). Direct sequencing of PCR products was attempted for several samples, however the majority of these failed to amplify. This is most likely due to the presence of methylation in the primer sequences preventing the attachment of primers, but indicates that DNA methylation patterns may vary across individuals and rearing environment. However, without further examination through cloning it is difficult to assess whether this methylation is random, non-random or simply a failure in the PCR protocol. Thus DNA methylation

does occur in *D. mojavensis* and this is likely due to the action of *Dnmt2*, however the functional significance of this methylation remains unresolved.

The results presented here demonstrate that the desaturase genes do not show fixed, population-specific genetic variation and that the variation present is likely to be to evolving through neutral processes. This study highlights the inherent problems in identifying candidate genes and narrowing down causative candidate mutations. Given the increasing number of techniques available for genome level studies it would seem prudent to use multiple and independent lines of investigation for the identification of candidate speciation genes. Only once a specific gene has been identified in more than one separate study can you have confidence that it worthy of further investigation. For example, QTL studies might improve resolution by narrowing down on a candidate genomic region, using an increased sample size and density of genetic markers. Alternatively, a QTL or genome-wide association study might complement a gene expression study in order to identify causative regulatory SNPs underlying a phenotype (Jones *et al*. 2012). Such intersection of different data types is likely to be an important future aspect of the on-going efforts to characterise the molecular basis of adaptation and speciation.

REFERENCES

Akiyama Y (1995) TFSEARCH: Searching transcription factor binding sites. http://www.rwcp.or.jp/papia.

Batterham P, Lovett JA, Starmer WT, Sullivan DT (1983) Differential regulation of duplicate alcohol-dehydrogenase genes in *Drosophila mojavensis*. *Developmental Biology*, **96**, 346-354.

Biron DG, Loxdale HD, Ponton F, *et al*. (2006) Population proteomics: An emerging discipline to study metapopulation ecology. *Proteomics*, **6**, 1712-1715.

Bousquet F, Ferveur JF (2012) *desat1*: A Swiss army knife for pheromonal communication and reproduction? *Fly (Austin)*, **6**, 102-107.

Bousquet F, Nojima T, Houot B, *et al*. (2012) Expression of a desaturase gene, *desat1*, in neural and nonneural tissues separately affects perception and emission of sex pheromones in *Drosophila*. *Proceedings of the National Academy of Sciences, USA*, **109**, 249-254.

Brazner JC, Etges WJ (1993) Pre-mating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. II. Effects of larval substrates on time to copulation, mate choice and mating propensity. *Evolutionary Ecology*, **7**, 605-624.

Coyne JA, Elwyn S (2006) Does the desaturase-2 locus in *Drosophila melanogaster* cause adaptation and sexual isolation? *Evolution*, **60**, 279-291.

Dallerac R, Labeur C, Jallon JM, *et al*. (2000) A Delta 9 desaturase gene with a different substrate specificity is responsible for the cuticular diene hydrocarbon polymorphism in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences, USA*, **97**, 9449-9454.

Dmitriev DA, Rakitov RA (2008) Decoding of superimposed traces produced by direct sequencing of heterozygous indels. *PLoS Computational Biology*, **4**, e1000113.

Drummond AJ, Ashton B, Buxton S, *et al*. (2011) Geneious v5. 4. Biomatter Ltd Auckland, New Zealand.

Etges WJ (1992) Premating isolation is determined by larval substrates in cactophilic *Drosophila mojavensis*. *Evolution*, **46**, 1945-1950.

Etges WJ, Ahrens MA (2001) Premating isolation is determined by larval-rearing substrates in cactophilic *Drosophila mojavensis*. V. Deep geographic variation in epicuticular hydrocarbons among isolated populations. *American Naturalist*, **158**, 585-598.

Etges WJ, de Oliveira CC, Gragg E, *et al*. (2007) Genetics of incipient speciation in *Drosophila mojavensis*. I. Male courtship song, mating success, and genotype x environment interactions. *Evolution*, **61**, 1106-1119.

Etges WJ, De Oliveira CC, Noor MAF, Ritchie MG (2010) Genetics of incipient speciation in *Drosophila mojavensis*. III. Life history divergence in allopatry and reproductive isolation. *Evolution*, **64**, 3549-3569.

Etges WJ, de Oliveira CC, Ritchie MG, Noor MAF (2009) Genetics of incipient speciation in *Drosophila mojavensis*. II. Host plants and mating status influence cuticular hydrocarbon QTL Expression and G × E Interactions. *Evolution*, **63**, 1712-1730.

Etges WJ, Jackson LL (2001) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. VI. Epicuticular hydrocarbon variation in *Drosophila mojavensis* cluster species. *Journal of Chemical Ecology*, **27**, 2125-2149.

Ewing AW (1979) Neuromuscular basis of courtship song in *Drosophila*: Role of the direct and axillary wing muscles. *Journal of Comparative Physiology*, **130**, 87-93.

Fang S, Takahashi A, Wu C-I (2002) A mutation in the promoter of *desaturase 2* is correlated with sexual isolation between *Drosophila* behavioral races. *Genetics*, **162**, 781-784.

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics*, **4**, 651-657.

Fellows DP, Heed WB (1972) Factors affecting host plant selection in desert-adapted cactophilic *Drosophila*. *Ecology*, **53**, 850-858.

Ferveur JF, Savarit F, Okane CJ, *et al*. (1997) Genetic feminization of pheromones and its behavioral consequences in *Drosophila* males. *Science*, **276**, 1555-1558.

Flot JF (2010) SEQPHASE: a web tool for interconverting phase input/output files and fasta sequence alignments. *Molecular Ecology Resources*, **10**, 162-166.

Gleason JM, Jallon JM, Rouault JD, Ritchie MG (2005) Quantitative trait loci for cuticular hydrocarbons associated with sexual isolation between *Drosophila simulans* and *Drosophila sechellia*. *Genetics*, **171**, 1789-1798.

Grillet M, Everaerts C, Houot B, *et al*. (2012) Incipient speciation in *Drosophila melanogaster* involves chemical signals. *Scientific Reports*, **2**, 224.

Heinemeyer T, Wingender E, Reuter I, *et al*. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Research*, **26**, 362-367.

Jiang MH, Zhang YH, Fei J, *et al*. (2010) Rapid quantification of DNA methylation by measuring relative peak heights in direct bisulfite-PCR sequencing traces. *Laboratory Investigation*, **90**, 282-290.

Jones FC, Grabherr MG, Chan YF, *et al*. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55-61.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059-3066.

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107-116.

Krauss V, Reuter G (2011) DNA methylation in *Drosophila*: a critical evaluation. *Progress in Molecular Biology and Translational Science*, **101**, 177-191.

Kunert N, Marhold J, Stanke J, Stach D, Lyko F (2003) A *Dnmt2*-like protein mediates DNA methylation in *Drosophila*. *Development*, **130**, 5083-5090.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451-1452.

Liu H, Wong L (2003) Data mining tools for biological sequences. *International Journal of Bioinformatics and Computational Biology*, **1**, 139-168.

Lyko F, Ramsahoye BH, Jaenisch R (2000) DNA methylation in *Drosophila melanogaster*. *Nature*, **408**, 538-540.

Lyko F, Ramsahoye BH, Kashevsky H, *et al*. (1999) Mammalian (cytosine-5) methyltransferases cause genomic DNA methylation and lethality in *Drosophila*. *Nature Genetics*, **23**, 363-366.

Machado CA, Matzkin LM, Reed LK, Markow TA (2007) Multilocus nuclear sequences reveal intra and interspecific relationships among chromosomally

polymorphic species of cactophilic *Drosophila*. *Molecular Ecology*, **16**, 3009-3024.

Markow TA (1991) Sexual isolation among populations of *Drosophila mojavensis*. *Evolution*, **45**, 1525-1529.

Matzkin LM (2004) Population genetics and geographic variation of alcohol dehydrogenase (*Adh*) paralogs and glucose-6-phosphate dehydrogenase (*G6pd*) in *Drosophila mojavensis*. *Molecular Biology and Evolution*, **21**, 276-285.

Matzkin LM, Eanes WF (2003) Sequence variation of alcohol dehydrogenase (*Adh*) paralogs in cactophilic *Drosophila*. *Genetics*, **163**, 181-194.

Mercot H, Defaye D, Capy P, Pla E, David JR (1994) Alcohol Tolerance, *Adh* Activity, and Ecological Niche of *Drosophila* Species. *Evolution*, **48**, 746-757.

Mettler LE (1963) *Drosophila mojavensis baja*, a new form in Mulleri complex. *Drosophila Information Service*, **38**, 421-430.

Nei M (1987) *Molecular evolutionary genetics* Columbia University Press, New York.

Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity*, **86**, 641-647.

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. *Nature Genetics*, **32**, 261-266.

Pfeiler E, Castrezana S, Reed L, Markow T (2009) Genetic, ecological and morphological differences among populations of the cactophilic *Drosophila mojavensis* from southwestern USA and northwestern Mexico, with descriptions of two new subspecies. *Journal of Natural History*, **43**, 923-938.

Rae PMM, Steele RE (1979) Absence of cytosine methylation at CCGG and GCGC sites in the rDNA coding regions and intervening sequences of *Drosophila* and the rDNA of other higher insects. *Nucleic Acids Research*, **6**, 2987-2995.

Reed LK, Markow TA (2004) Early events in speciation: polymorphism for hybrid male sterility in *Drosophila*. *Proceedings of the National Academy of Sciences, USA*, **101**, 9009.

Reed LK, Nyboer M, Markow TA (2007) Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. *Molecular Ecology*, **16**, 1007-1022.

Richards EJ (2006) Inherited epigenetic variation - revisiting soft inheritance. *Nature Reviews Genetics*, **7**, 395-401.

Richards EJ (2008) Population epigenetics. *Current Opinion in Genetics & Development*, **18**, 221-226.

Ritchie MG, Noor MAF (2004) Evolutionary genetics - Gene replacement and the genetics of speciation. *Heredity*, **93**, 1-2.

Rockman MV (2012) The QTN program and the alleles that matter for evolution: All that's gold does not glitter. *Evolution*, **66**, 1-17.

Ross C, Markow T (2006) Microsatellite variation among diverging populations of *Drosophila mojavensis*. *Journal of Evolutionary Biology*, **19**, 1691-1700.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365-386.

Ruiz A, Heed W, Wasserman M (1990) Evolution of the mojavensis cluster of cactophilic *Drosophila* with descriptions of two new species. *Journal of Heredity*, **81**, 30-42.

Schaefer M, Pollex T, Hanna K, *et al*. (2010) RNA methylation by *Dnmt2* protects transfer RNAs against stress-induced cleavage. *Genes & Development*, **24**, 1590-1595.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, **141**, 413.

Stennett MD, Etges WJ (1997) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. III. Epicuticular hydrocarbon variation is determined by use of different host plants in *Drosophila mojavensis* and *Drosophila arizonae*. *Journal of Chemical Ecology*, **23**, 2803-2824.

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, **76**, 449-462.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978-989.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158-170.

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

Takahashi A, Tsaur SC, Coyne JA, Wu C-I (2001) The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences, USA*, **98**, 3920-3925.

Toolson EC, Markow TA, Jackson LL, Howard RW (1990) Epicuticular hydrocarbon composition of wild and laboratory-reared *Drosophila mojavensis* patterson and crow (diptera, drosophilidae). *Annals of the Entomological Society of America*, **83**, 1165-1176.

CHAPTER 3:

MODEL-BASED COMPARISONS OF PHYLOGEOGRAPHIC SCENARIOS RESOLVE THE INTRASPECIFIC DIVERGENCE OF CACTOPHILIC *DROSOPHILA MOJAVENSIS*

MATERIAL FROM THIS CHAPTER PUBLISHED AS:

Smith G, Lohse K, Etges WJ, Ritchie MG (2012) Model-based comparisons of phylogeographic scenarios resolve the intraspecific divergence of cactophilic *Drosophila mojavensis*. *Molecular Ecology*, **13**, 3293-3307.

AUTHOR CONTRIBUTIONS:

GS and MGR conceived the project. WJE provided flies and experimental advice. GS performed all sequencing and IMa2 analyses, KL performed replicate IMa2 analyses and BEAST analyses. GS wrote the paper with contributions from all co-authors.

ABSTRACT

The cactophilic fly *Drosophila mojavensis* exhibits considerable intraspecific genetic structure across allopatric geographic regions and shows associations with different host cactus species across its range. The divergence between these populations has been studied for more than 60 years, yet their exact historical relationships have not been resolved. We analysed sequence data from 15 intronic X-linked loci across populations from Baja California, mainland Sonora-Arizona and Mojave Desert regions under an isolation-with-migration model to assess multiple scenarios of divergence. We also compared the results with a pre-existing sequence dataset of 8 autosomal loci. We derived a population tree with Baja California placed at its base and link their isolation to Pleistocene climatic oscillations. Our estimates suggest the Baja California population diverged from an ancestral Mojave/Mainland group around 230-270 Kya, while the split between the Mojave Desert and Mainland populations occurred one glacial cycle later, 117-135 Kya years ago. Although we found these three populations to be effectively allopatric, model ranking could not rule out the possibility of a low-level of gene flow between two of them. Finally, the Mojave Desert population showed a small effective population size, consistent with a historical population bottleneck. We show that model-based inference from multiple loci can provide accurate information on the historical relationships of closely related groups allowing us to set into historical context a classic system of incipient ecological speciation.

INTRODUCTION

The role of adaptation and ecological specialisation in speciation has enjoyed a recent resurgence of interest (Rundle & Nosil 2005; Schluter 2001). Although natural selection has always been seen as key to the biology of speciation (Sobel *et al*. 2010) recent studies and models make more explicit links between ecological adaptation and reproductive isolation. For example, if ecological selection either directly or indirectly influences traits involved in assortative mating, natural and sexual selection may act in partnership to drive rapid speciation (Funk *et al*. 2006; Maan & Seehausen 2011; Van Doorn *et al*. 2009; Weissing *et al*. 2011). *Drosophila* is often seen as a key system for understanding the evolution of reproductive incompatibilities, but few species have been studied in the context of ecological speciation (Mallet 2006). Some species are ecological specialists, including some mycophagous *Drosophila* (Jaenike 1990), numerous Hawaiian drosophilids (Kambysellis *et al*. 1995; Magnacca *et al*. 2008), and the island endemic *D. sechellia* that is uniquely adapted to volatile toxic components of *Morinda citrifolia* fruit (Farine *et al*. 1996; R'kha *et al*. 1991). The adaptation of some *Drosophila* species to live and breed on different species of fermenting cactus provides a potential example of ecological speciation as adaptation to different species of cactus can influence reproductive isolation (Etges *et al*. 2010). However the relative contribution of host plant specialization and population history to the population divergence of cactophilic species is not yet clear (Machado *et al*. 2007). Here we investigate the history of three populations of a well-studied cactophilic fly *Drosophila mojavensis* using molecular data.

D. mojavensis is a cactophilic member of the repleta group, within the *Drosophila* subgenus. This species is distributed throughout the Sonoran and Mojave Deserts and adjacent arid lands in Baja California, southern California, north-western

Mexico and Arizona, USA where it lives in a number of host cacti (Fellows and Heed 1972; Ruiz *et al*. 1990) (Fig. 3.1).

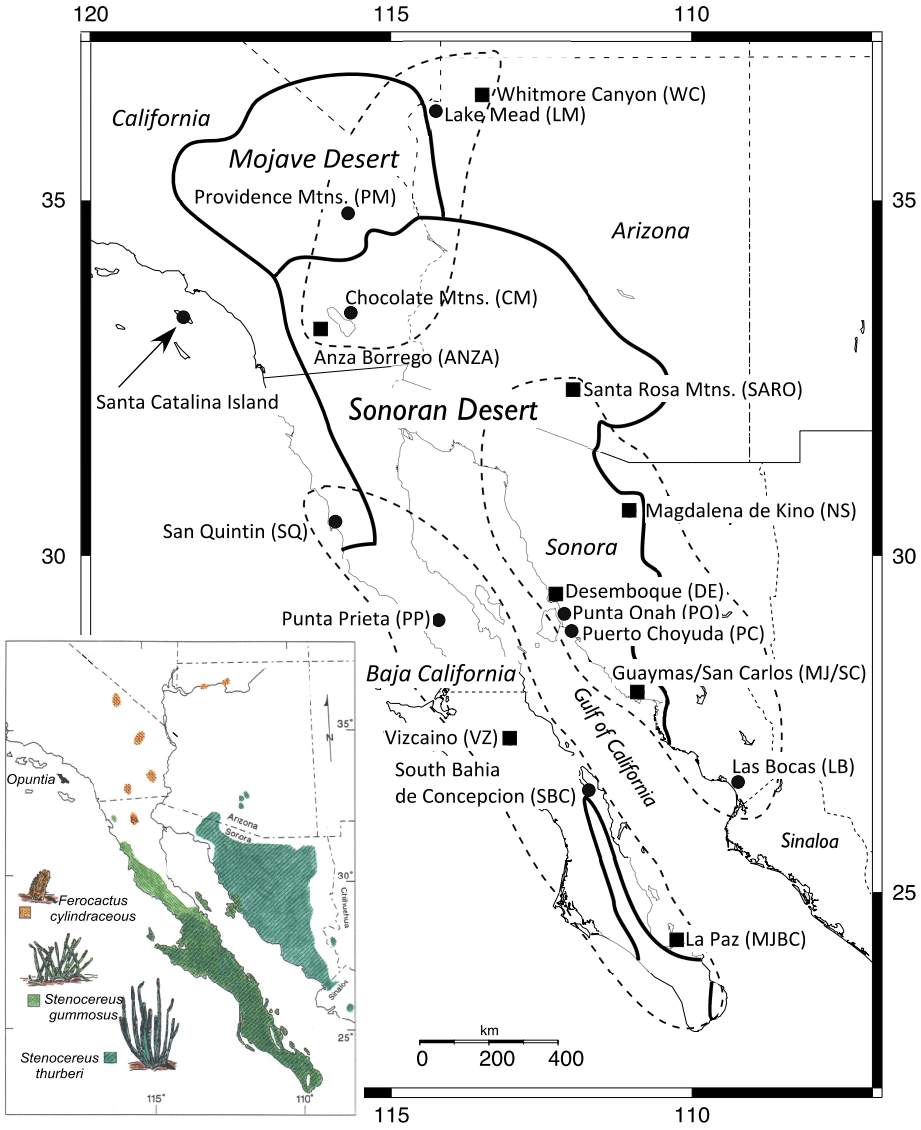


Fig. 3.1 Map of the Southwestern USA and Northwestern Mexico with approximate boundaries of the Sonoran and Mojave Deserts (black lines). Sample site location names and abbreviations are shown; darkened circles indicate sites described in this study, and darkened squares sites from Machado *et al*. (2007). Dashed lines denote the ranges of the populations used in this study; the Baja California, Mojave Desert and Mainland populations. Santa Catalina Island is indicated, but not included in this study. The inset map shows the ranges of the four major host cacti used by *D. mojavensis*; *Ferocactus cylindraceous* in the Mojave Desert, *Stenocereus gummosus* in Baja California, the islands on the Gulf of California, and a small patch in coastal Sonora, and *S. thurberi* in southern Baja California, and mainland Mexico and Arizona.

Populations of *D. mojavensis* are considered oligophagic, typically only using one host cactus species in different parts of the species range, with occasional use of other host species. The Baja California population uses agria cactus, *Stenocereus gummosus*, while *D. mojavensis* in north-western Mexico (principally in Sonora and Sinaloa) and Southern Arizona use organ pipe cactus *S. thurberi,* although they sometimes share sina cactus, *S. alamosensis*, with sister species *D. arizonae* in Sonora and Sinaloa. In southern California the Mojave Desert population uses barrel cactus *Ferocactus cylindraceous* and on Santa Catalina Island near Los Angeles, California flies use prickly pear cactus species including *Opuntia demissa* and *O. littoralis*. *D. mojavensis* exhibits more population structure than *D. arizonae*, which has a far wider distribution ranging from Arizona and New Mexico, USA to Chiapas in southern Mexico (Heed 1978; Ruiz & Heed 1988; Wasserman 1992) and is associated with different host cacti. Significant levels of sexual isolation have been reported between mainland Mexico and Baja California populations of *D. mojavensis* which has lead to them being described as incipient species (Etges *et al*. 2007; Etges *et al*. 2010; Etges *et al*. 2009; Etges *et al*. 2008; Markow 1991; Pfeiler *et al*. 2009; Zouros & d'Entremont 1980). However, detailed studies have demonstrated this isolation to be influenced by the species of cactus on which flies develop (Brazner & Etges 1993; Etges 1992; Stennett & Etges 1997). Common garden experiments have shown that rearing Baja or mainland Mexico flies reciprocally on organ pipe and agria cactus influences not only survival and life history but also courtship behaviours and assortative mating. Further, there are extensive genotype-by-environment interactions for Quantitative Trait Loci (QTLs) influencing life history and behavioural traits (Etges *et al*. 2007; Etges *et al*. 2009). Thus the *D. mojavensis* system shows

significant evidence of ecological specialization, influencing a range of traits associated with adaptation and differentiation.

Although significant genetic structure is known within *D. mojavensis* (Machado *et al*. 2007; Ross & Markow 2006), the order of population divergence remains unresolved, and the geographic region from which the species is thought to have originated has been disputed. Phylogenetic analysis of second chromosome gene arrangements provided strong evidence that the Baja California peninsula is the centre of genetic diversity because of a rare ancestral chromosome in central Baja Californian flies (Johnson 1980; Wasserman 1992). Sequence variation in *ADH* enzyme loci, as well as glucose-6-phosphate dehydrogenase (*G6pd*) also support a Baja California origination of *D. mojavensis* (Matzkin 2004; Matzkin & Eanes 2003). In contrast Reed *et al*. (2007) suggested the Mainland region to be the centre of diversity, concluding a later colonization of the Mojave Desert and Baja California regions. Machado *et al*. (2007) demonstrated that the Santa Catalina Island and Mojave Desert populations group together on a separate lineage to the Baja and Mainland populations. Ross and Markow (2006) also concluded that the Baja and Mainland populations were more closely related to one another than the other two groups. The only attempt at dating population differentiation came from Reed *et al*. (2007) using mitochondrial DNA, estimating population differentiation to have occurred around 270-690 Kya. Thus, after more than 60 years of study, the order and timing of population divergence of this classic system have yet to be resolved. The identification of the 'ancestral' group is necessarily based on analysis of contemporary populations. Hereafter, when we refer to the 'ancestral' population (or populations) we explicitly mean an inferred ancestral group from contemporary population genetic data.

It is unknown to what extent the differentiation of populations in these regions has been influenced by gene flow rather than shared ancestral polymorphism. QTL analysis of Baja California and Mainland populations revealed high levels of shared polymorphisms, genotype-by-environment interactions and transgressive genetic variation, perhaps suggesting that gene flow between regions is on-going (Etges *et al*. 2007; Etges *et al*. 2009). Distinguishing isolation from migration is a key challenge in population genetic inference (Hey & Nielsen 2004). Traditional analyses of the relationships between populations often involve allelic markers or mtDNA sequences and have limited power to distinguish between shared ancestral polymorphism and gene flow (Zhang & Hewitt 2003). In contrast, joint analysis of sequence data from multiple nuclear loci in a model-based coalescent framework allows for accurate quantification of recent population histories (Garrick *et al*. 2010; Hey 2005). Such a framework also has the advantage of not relying on large numbers of individuals, but rather increases its power by sampling across many loci (Hey & Nielsen 2004; Lohse *et al*. 2010; Wang & Hey 2010).

Here we use multiple intronic loci to infer the order in which the three main populations in the *D. mojavensis* species range (i.e. Baja California, Mojave Desert, and Mainland) diverged, and estimate the demographic parameters associated with their history. We used likelihood-based model selection for each pairwise comparison of populations to infer the most likely scenario of divergence and obtained estimates of demographic parameters based on these models. These parameters were then used to inform a full, three-population model. Further, to assess these results, we ran analogous analyses on an independent, autosomal dataset from Machado *et al* (2007), comprised of 7 Kb of sequence over 8 loci each including both coding and non-coding regions. We show that the use of multiple loci in a model-based coalescent framework

can distinguish isolation from migration, be informative of demographic parameters and provide information on the evolutionary relationships of recently diverged populations.

METHODS

Primer design and sequencing

Individuals were sampled from across the three major geographic regions (Fig. 3.1). DNA was extracted from whole flies using cell lysis solution (0.1M EDTA, 0.2M Tris and 1% SDS) and incubated with Proteinase K and RNase A. Fifteen primer sets were designed to amplify intronic regions of the X-chromosome (Supplementary Table S3.4). Loci were spread evenly across the chromosome with an average of 1.5Mb between them. Large (700-900bp) introns were randomly sampled from coding genes. Extracted DNA was amplified by PCR, and cleaned with a Qiagen MinElute PCR Purification Kit. Sequences were obtained from males only to allow for direct sequencing without cloning or computational phasing. Sequencing was carried out on an ABI 3730 instrument following a BigDye reaction (sequences available in GenBank under accession numbers JQ405359-JQ405528). Sequence length varied from 500-600bp (after the trimming described below). In total, twelve individuals were sequenced for fifteen loci (Baja N=4, Mainland N=3 and Mojave N=5). Sequences were aligned using MAFFT (Katoh *et al*. 2002) and all indels were excluded from further analyses. Each chromatogram and alignment was checked by eye for accuracy and low quality sequences were re-sequenced.

We also reanalysed data from Machado *et al*. (2007) obtained from Genbank, which included 8 coding and non-coding nuclear loci from chromosomes 2, 3, 4 and 5 for 15 individuals from the three populations (see Fig. 3.1 and Table 3.1). After

testing for recombination and trimming alignments into non-recombining blocks, these autosomal sequences were an average of 860bp in length. Two X loci in the original dataset of 10 loci were excluded from our analyses.

Table 3.1 Sample sites, including the location under which each population was grouped for analyses, the number of individuals sampled from each population and the source of each sample. Collection year is given in brackets for samples sequenced in this analysis (See Machado *et al*. 2007 for more details).

Population	Sample site	Number Individuals	Source
Baja California	San Quintin (SQ)	2	Lab (2008)
	Punta Prieta (PP)	1	Lab (2008)
	South Bahiá Conceptión (SBC)	1	Lab (1996)
	Vizcaino (VZ)[m]	2	Lab
	La Paz (MJBC)[m]	1	Lab
Mainland	Punta Onah (PO)	1	Wild (2009)
	Puerto Choyuda (PC)	1	Wild (2009)
	Las Bocas (LB)	1	Wild (2009)
	Santa Rosa Mountains (SARO)[m]	1	Lab
	Magdalena de Kino (NS)[m]	3	Lab
	Desemboque (DE)[m]	1	Lab
	Guaymas/San Carlos (MJ/SC)[m]	3	Lab
Mojave Desert	Lake Mead (LM)	1	Lab (2010)
	Providence Mountains (PM)	1	Lab (1996)
	Chocolate Mountains (CM)	3	Lab (1988)
	Whitmore Canyon (WC)[m]	2	Lab
	Anza-Borrego Desert (ANZA)[m]	2	Lab

[m] , denotes data taken from Machado *et al*. (2007).

Testing assumptions of the isolation-with-migration model

The isolation-with-migration (IM) model is a non-equilibrium description of population relationships that allows populations to be related by divergence from a common ancestral population and through migration after divergence (Hey & Nielsen 2004). For two populations the model includes six parameters; the divergence time t, two migration rates between descendent populations M in either direction and the effective population sizes N_e of the ancestral and both descendent populations. IMa2 parameters are scaled relative to μ, the geometric mean of the (per generation)

mutation rate across loci. Migration is measured as $m = M/\mu$, scaled effective population sizes are defined as $\theta = 4N_e\mu$ and divergence time as $T = t/u\ g$, where M is the rate of migration per generation, t and T denote population divergence measured in numbers of generations and years respectively and g is the generation time (in years). The model makes the standard population genetic assumptions of random mating within populations, large effective size and selective neutrality. We used IMa2 (Hey 2010; Hey & Nielsen 2004) to estimate parameters under this model. IMa2 uses Markov Chain Monte Carlo sampling (MCMC) in a Bayesian framework to estimate parameters under the IM model and makes the additional assumptions of no recombination within, but free recombination between loci, and no other groups connected to the study populations via gene flow. All assumptions are reasonable in our study. First, there is little evidence of population sub-structure within each of the three *D. mojavensis* populations, except perhaps for Baja California (Etges *et al*. 1999; Ross & Markow 2006). Second, given the physical distance between loci, ignoring physical linkage between them is justified. Finally, we tested for recombination within loci using Recombitest (Piganeau *et al*. 2004). Two loci (3 and 6) showed evidence of recombination based on two different measures of linkage disequilibrium (r^2 and $|D'|$) correlated with distance between sites, so both were trimmed to the largest block compatible with the assumption of no-recombination (Piganeau & Eyre-Walker 2004). We used DNAsp version 5.10.01 (Librado & Rozas 2009) to estimate genetic diversity and appropriate nucleotide substitution models were identified for each locus using jmodeltest (Posada 2008). The HKY model provided the best fit for all X-chromosome loci and several of the Machado *et al*. (2007) loci. Since IMa2 only supports the infinite sites and HKY models for DNA

sequence data, we chose the more complex HKY model as being more appropriate for both datasets.

Coalescent analyses using IMa2

Given the large number of possible parameter combinations and the need to specify the order of divergence in a three-population model (unknown in *D. mojavensis*), we first ran separate IMa2 analyses for all three pairwise comparisons of *D. mojavensis* populations followed by three-population models using the inferred topology and reduced (most likely) parameter set. While an exhaustive comparison of all possible models is feasible for two-population IM scenarios, for three or more populations the sheer number of models makes this approach impractical. Consider the full three-population IM model, which has 8 migration and 5 effective population size parameters. There are 2^8 ways to simplify this model by setting one (or more) migration rates to zero and 2^4 ways to set one (or more) pairs of migration rates to be symmetrical. Likewise there are 2^4 model simplifications that involve setting one (or more) population sizes to be equal to ancestral population sizes. Thus, there are at least $(16 + 256) \times 16 = 4352$ nested models for a fixed population tree topology. Note that this excludes model simplifications that involve both symmetric migration rates between some pairs of populations and zero gene flow between others, so the number of all possible models is substantially larger. In other words to explore the full space of three-population IM scenarios, one would need to compare over 13,000 models, which is not feasible. Instead we adopted a bottom-up strategy, which consisted of comparing nested models for each of the three pairwise comparisons of populations and using these to specify a three-population model. Our rationale was that if gene flow has been limited the omitted population would not contribute to the analysis and

thus the divergence estimates obtained in these pairwise runs would be informative about the order of divergence. Conversely, if population relationships were dominated by gene flow, this should be detectable as high migration rate estimates.

Mutation rates can be specified in IMa2 to convert scaled (in units of N_e generations) parameters into absolute values. For *D. melanogaster* a direct, lab-based estimate of the genome-wide mutation rate is available (Keightley *et al*. 2009) which avoids potentially unreliable indirect rate calibrations (Pulquerio & Nichols 2007). Keightley *et al*. (2009) measured the mutation rate of three mutation accumulation lines over an average of 262 generations and determined a single nucleotide mutation rate of 3.5×10^{-9} per site per generation. We used this direct mutation rate measure and also incorporated the 95% CI of this estimate (2.96×10^{-9}, 4×10^{-9}) into our analyses. Although we used a genome-wide estimate applied to introns, little is known about intron function and evolution. Thus we believe a direct mutation rate to be broadly appropriate. We multiplied the direct mutation rate by the length of each alignment (assuming there was no constraint on intron divergence) to convert to per locus mutation rates. In the Machado *et al*. (2007) dataset, non-synonymous sites were excluded from this calibration. The inheritance scalar for the X chromosome data was set to 0.75. We assumed 6 generations per year for *D. mojavensis* consistent with lab generation times, our knowledge of the seasonal phenology of breeding site availability (Etges, pers. observation) and previous estimates (Matzkin & Eanes 2003).

We performed extensive runs for all 3 pairwise comparisons to optimise the MCMC settings and identify appropriate prior bounds. We used a geometric heating scheme with 40 chains (a = 0.975, b = 0.7) and upper prior bounds for parameters were: $t = 15$, $N_e = 37.5$ and $m = 0.7$. For each pairwise comparison, we initially ran IMa2 for the full model using the above settings, a burn-in of 10,000 generations and

sampling 100,000 genealogies. Separate analyses were performed for the X and the Machado *et al*. (2007) datasets. To test for convergence, all runs were repeated with a different random number seed.

We used genealogies sampled in the initial MCMC runs to estimate the joint likelihoods of all simpler models nested within the full, six-parameter model (using the L-mode in IMa2). There are 25 ways of simplifying the full model (Table 3.4), either by setting migration parameters to zero or by combining parameters. For each nested model we compared the joint log likelihood (LogL) to that of the full model to find the best model(s) given the data. We adopted the method of Carstens *et al*. (2009) and ranked models using Akaike Information Criterion (AIC) scores (Akaike 1973), defined as AIC = 2k – LogL. This penalises the joint log likelihood (LogL) of each model by its complexity, where k is the number of model parameters. Although we were interested in identifying the best model, ranking a set of likely models allowed us to move beyond accept/reject hypothesis testing and investigate the relationships between the three *D. mojavensis* populations in more detail (Carstens *et al*. 2009). Following Carstens *et al*. (2009) we also calculated two related information theoretical statistics to facilitate model comparisons: Akaike weights (ω_i) i.e. the normalized relative likelihoods of the model, and the evidence ratio ($E_{min/I} = \omega_{min} / \omega_i$), which compares each model to the best model and provides an objective measure of model support. We considered an evidence ratio of <10 to be moderate support for a model relative to the top model (Anderson & Burnham 2002).

Having identified a set of plausible models, we ran a second IMa2 analysis for each pairwise comparison under the best-supported simplified IM model to obtain final parameter estimates. Finally, we ran a three-population model given the most likely population tree topology as identified from the divergence times in the

simplified pairwise runs, MCMC and prior settings were set as described above for the two-populations models. Model selection results from the pairwise comparisons were used to simplify the three-population model, by removing unsupported migration parameters.

To independently examine the tree topology we also analysed our data in *BEAST (Drummond & Rambaut 2007), which does not require a pre-defined tree topology but is limited to strict divergence models. Priors of the *BEAST analysis were set to be identical to those used in the IMa analyses whenever possible. We assumed a HKY nucleotide substitution model and a fixed per site mutation rate across loci as derived by Keightley *et al*. (2009). We ran six replicate *BEAST runs each with 10 million generations and a burn-in of 1 million, sampling 10,000 population trees per run. The posterior samples of trees from the six replicates were combined to produce a maximum clade credibility tree (assuming a strict molecular clock), including divergence times and clade support values. Trees were visualised in FigTree v1.3.1 (Rambaut 2010).

RESULTS

A total of 9Kb from fifteen X chromosome loci was sequenced from 12 individuals across three geographically isolated populations of *D. mojavensis*; Baja California, Mojave Desert and Mainland (Table 3.1). A population on Santa Catalina Island (Fig. 3.1) was not included. The total number of polymorphic sites (285) and mean per site nucleotide diversity for all loci (0.014) indicated high overall genetic variation, comparable to that seen in African *Drosophila melanogaster* (π=0.0114; Ometto *et al*. 2005; Table 3.2). The Mojave Desert population (Supplementary Table S3.1) had a lower diversity (π=0.0022) than the Baja California and Mainland populations

(π=0.017 and π=0.019 respectively) in agreement with previous studies (Ross & Markow 2006). Diversity in the Machado *et al*. (2007) dataset (π=0.0079) was almost half that seen in the X data (Supplementary Table S3.1).

Table 3.2 Summary statistics of each locus sampled in the X chromosome dataset

Locus	Gene identifier	*L*	*S*	*HD*	π_S	θ
1	GI16459	835	15	0.848	0.0161	0.0171
2	GI11083	652	7	0.833	0.0044	0.0054
3	GI11153	584	31	0.945	0.0481	0.0516
4	GI21637	652	19	0.848	0.0172	0.0184
5	GI21676	676	20	0.818	0.0114	0.0153
6	GI21740	674	13	0.891	0.0094	0.0104
7	GI21817	556	19	0.933	0.0243	0.0261
8	GI15371	587	25	0.844	0.0064	0.0080
9	GI15520	665	7	0.891	0.0119	0.0157
10	GI15595	358	11	0.956	0.0118	0.0133
11	GI15666	642	16	0.917	0.0166	0.0176
12	GI15707	524	10	0.618	0.0067	0.0115
13	GI15862	644	10	0.909	0.0115	0.0146
14	GI15934	472	48	0.867	0.0072	0.0092
15	GI15975	640	34	0.867	0.0126	0.0136
	Total	9161	285			
	Mean	610.7	19			

L, length of locus alignment; S, number of polymorphic sites; HD, haplotype diversity; π_S, per site nucleotide diversity (Nei, 1987); θ, from Watterson (1975).

IMa2 and pairwise model selection

ESS values, visual examination of parameter trend plots and near identical results of replicate MCMC runs indicated good mixing and convergence of the Markov chains. In all three pairwise comparisons the best-supported model (with the lowest AIC score) contained no migration (Table 3.3, see Table 3.4 for model descriptions). However, several models with low rates of migration in each comparison were found to be moderately supported (below the evidence ratio cut off of 10), but with little information on the direction of the migration. Although a strict divergence model (Model 5) had the highest support for the Mojave/Baja comparison (Table 3.3c), its

evidence ratio was almost identical to that of the next model (Model 2), which included a low rate of symmetric migration (1.06:1 odds in favour of the simple divergence model). Thus, although our estimates of gene flow were very low we could not rule out migration between Mojave and Baja and therefore considered both Model 2 and Model 5 for final parameter estimates for the Mojave/Baja comparison, and zero migration (Model 5) for the remaining two comparisons.

Table 3.3 Summary of model selection results for the three X chromosome comparisons. Shown are models with moderate support ($E_{min/i} < 10$) for (a) Baja vs Mainland, (b) Mainland vs Mojave and (c) Mojave vs Baja

(a)	Model number	log(P)	K	θ_0	θ_1	θ_2	$m0>1$	$m1>0$	AIC	Δ_i	ω_i
	15	-3.684	2	6.738	37.5	[6.7379]	[0.00000]	[0.00000]	20328.938	0.000	0.266219
	14	-3.322	3	5.610	37.5	[5.6104]	0.162	[0.00000]	20330.214	1.276	0.140656
	5	-3.487	3	7.861	37.5	6.010	[0.00000]	[0.00000]	20330.544	1.606	0.119262
	12	-3.495	3	6.283	37.5	[6.2833]	0.032	[0.0319]	20330.560	1.622	0.118311
	13	-3.552	3	6.474	37.5	[6.4739]	[0.00000]	0.043	20330.674	1.736	0.111756
	4	-3.276	4	6.267	37.5	5.338	0.134	[0.00000]	20332.122	3.184	0.054180
	3	-3.304	4	7.717	37.5	5.274	[0.00000]	0.066	20332.178	3.240	0.052684
	2	-3.313	4	7.242	37.5	5.257	0.035	[0.0347]	20332.196	3.258	0.052212
	11	-3.322	4	5.610	37.5	[5.6104]	0.162	0.000	20332.214	3.276	0.051745

(b)	Model Number	log(P)	K	θ_0	θ_1	θ_2	$m0>1$	$m1>0$	AIC	Δ_i	ω_i
	20	-0.848	2	0.716	10.650	[10.6501]	[0.00000]	[0.00000]	20839.778	0.000	0.308610
	5	-0.742	3	0.722	8.299	11.158	[0.00000]	[0.00000]	20841.566	1.787	0.126264
	18	-0.780	3	0.727	10.680	[10.6798]	[0.00000]	0.024	20841.643	1.864	0.121508
	17	-0.814	3	0.704	10.675	[10.6753]	0.016	[0.0163]	20841.710	1.931	0.117505
	19	-0.848	3	0.716	10.650	[10.6501]	0.000	[0.00000]	20841.778	2.000	0.113531
	3	-0.631	4	0.736	8.024	11.352	[0.00000]	0.031	20843.344	3.566	0.051888
	2	-0.687	4	0.709	8.192	11.285	0.020	[0.0195]	20843.456	3.678	0.049062
	4	-0.742	4	0.722	8.299	11.158	0.000	[0.00000]	20843.566	3.787	0.046450
	16	-0.780	4	0.727	10.680	[10.6798]	0.000	0.024	20843.643	3.864	0.044700

(c)	Model Number	log(P)	K	θ_0	θ_1	θ_2	$m0>1$	$m1>0$	AIC	Δ_i	ω_i
	5	-0.829	3	0.821	37.5	7.455	[0.00000]	[0.00000]	20004.872	0.000	0.277306
	2	0.110	4	0.709	37.5	5.855	0.051	[0.051]	20004.994	0.122	0.260896
	4	-0.466	4	0.704	37.5	6.447	0.063	[0.00000]	20006.146	1.274	0.146690
	3	-0.533	4	0.862	37.5	7.137	[0.00000]	0.038	20006.281	1.409	0.137115
	1	0.138	5	0.696	37.5	5.833	0.064	0.043	20006.938	2.066	0.098694

Model number, see Table 3.4; K, number of model parameters; *m*, migration '>', direction of migration backwards in time; [], number within brackets fixed; Δ_i, difference in AIC; ω_i, Aikaike weights; $E_{\text{min/I}}$, Evidence ratio.

(a) θ_0, Mainland effective population size, θ_1, Baja *Ne*; θ_2, ancestral N_e

(b) θ_0, Mojave effective population size, θ_1, Mainland N_e; θ_2, ancestral N_e

(c) θ_0, Mojave effective population size, θ_1, Baja *Ne*; θ_2, ancestral N_e

Model number	Model Description
1	FULL
2	m all equal
3	m 0>1 is zero
4	m 1>0 is zero
5	m = zero
6	$\theta_0=\theta_1$
7	$\theta_0=\theta_1$, m all equal
8	$\theta_0=\theta_1$, m 0>1 is zero
9	$\theta_0=\theta_1$, m 1>0 is zero
10	$\theta_0=\theta_1$, m = zero
11	$\theta_0=\theta_2$
12	$\theta_0=\theta_2$, m all equal
13	$\theta_0=\theta_2$, m 0>1 is zero
14	$\theta_0=\theta_2$, m 1>0 is zero
15	$\theta_0=\theta_2$, m = zero
16	$\theta_1=\theta_2$
17	$\theta_1=\theta_2$, m all equal
18	$\theta_1=\theta_2$, m 0>1 is zero
19	$\theta_1=\theta_2$, m 1>0 is zero
20	$\theta_1=\theta_2$, m = zero
21	θ all equal
22	θ all equal, m all equal
23	θ all equal, m 0>1 is zero
24	θ all equal, m 1>0 is zero
25	θ all equal, m = zero

Table 3.4 Model selection (L mode) model descriptions for all 25 possible nested models within the full two-population IM model.

For each pairwise comparison we carried out likelihood ratio tests to test each model against the full model (Model 1; Tables S3.6-3.8). Significance at p = 0.05 was assessed both with and without Bonferroni correction. Using an evidence ratio cut-off of 10 for ranked models generally included comparisons that were non-significant at 5% (i.e. the models cannot be rejected). Applying a correction for multiple testing reduced the number of models that could be rejected (Tables S3.6-3.8).

Simplified model results

Our parameter estimates from the top selected models (Table 3.3) suggest that divergence of *D. mojavensis* populations occurred several hundred thousand years ago, with very little subsequent gene flow. Divergence of the Mojave and Mainland populations (117 Kya) postdates divergence between both Mojave/Baja (177 Kya no migration, 226 Kya with equal migration) and Mainland/Baja (254 Kya) (Table 3.5a). Although 95% highest posterior density (HDP) intervals of all three divergence time estimates overlap, the similar divergence time estimates for the Mojave/Baja (when migration was in the model) and the Mainland/Baja splits suggest that the most recent common ancestor of all three populations divided into a Baja population and a Mojave/Mainland ancestral population first, with the divergence of the Mojave/Mainland ancestor into contemporary Mojave and Mainland populations coming later (Table 3.5a and Fig. 3.2).

Table 3.5 Posterior high points and 95% HPD (in brackets) of model parameters estimated under the most parsimonious models for (a) two-population models for both the X chromosome and Machado *et al*. (2007) datasets, where 'Mojave/Baja (migration)' is the comparison under Model 2 of equal migration and (b) three-population models.

(a)

	X chromosome				**Machado et al.**		
Parameter	Baja vs. Mainland	Mojave vs. Mainland	Mojave vs. Baja	Mojave vs. Baja (migration)	Baja vs. Mainland	Mojave vs. Mainland	Mojave vs. Baja
Divergence time (95% HPD)	254K (171-336K)	117K (57-203K)	177K (102-259K)	226K (108-383K)	182K (108-267K)	230K (130-363K)	284K (134-420K)
Mainland N_e (95% HDP)	1.65M (817K-4M)	1.7M (1.5M-3.7M)	-	-	2.7M (1.6-2.4M)	2.1M (1.3-3.6M)	-
Baja N_e (95% HPD)	*	-	*	*	*	-	1.9 M (843K-5.7M)
Mojave N_e (95% HPD)	-	149K (68-278K)	181K (93-318K)	149K (68-278)	-	330K (156-662K)	406K (191-822K)
Ancestral N_e (95%HPD)	1.2M (615K-2.1M)	2.3M (1.5-3.6M)	1.5M (801K-2.5M)	1.2M (431K-2.3M)	1.1M (551K-2M)	1.1M (461K-2.1M)	1.1M (281K-2.1M)

(b)

Parameter	No migration	Equal migration
Divergence time$_0$ (95% HPD)	123K (59-211K)	135K (63-231K)
Divergence time$_1$ (95% HPD)	229K (117-290K)	270K (195-383K)
Mainland N_e (95% HDP)	1.25M (487K-5.3M)	1.34M (519K-5.8M)
Baja N_e (95% HPD)	*	*
Mojave N_e (95% HPD)	149K (68-285K)	109K (44-237K)
Ancestral N_{e0} (95%HPD)	1.85M (648K-7.4M)	978K (221K-6.3M)
Ancestral N_{e1} (95%HPD)	800K (487K-1.3M)	624K (262K-1.1M)
$m_{\text{Mojave>Baja}}$	-	0.08
$m_{\text{Baja>Mojave}}$	-	0.00
$m_{\text{Ancest>Baja}}$	-	0.00
$m_{\text{Baja>Ancest}}$	-	0.7

Divergence time$_0$, Mojave/Mainland divergence time; Divergence time$_1$, Baja divergence time from Mojave/Mainland ancestor; Ancestral N_{e0}, size of Mojave/Mainland ancestor; Ancestral N_{e1}, ancestor of all; m, migration '>' indicates direction backwards in time; * Indicates a parameter that was not estimated. Divergence times are in years and effective population sizes in numbers of individuals.

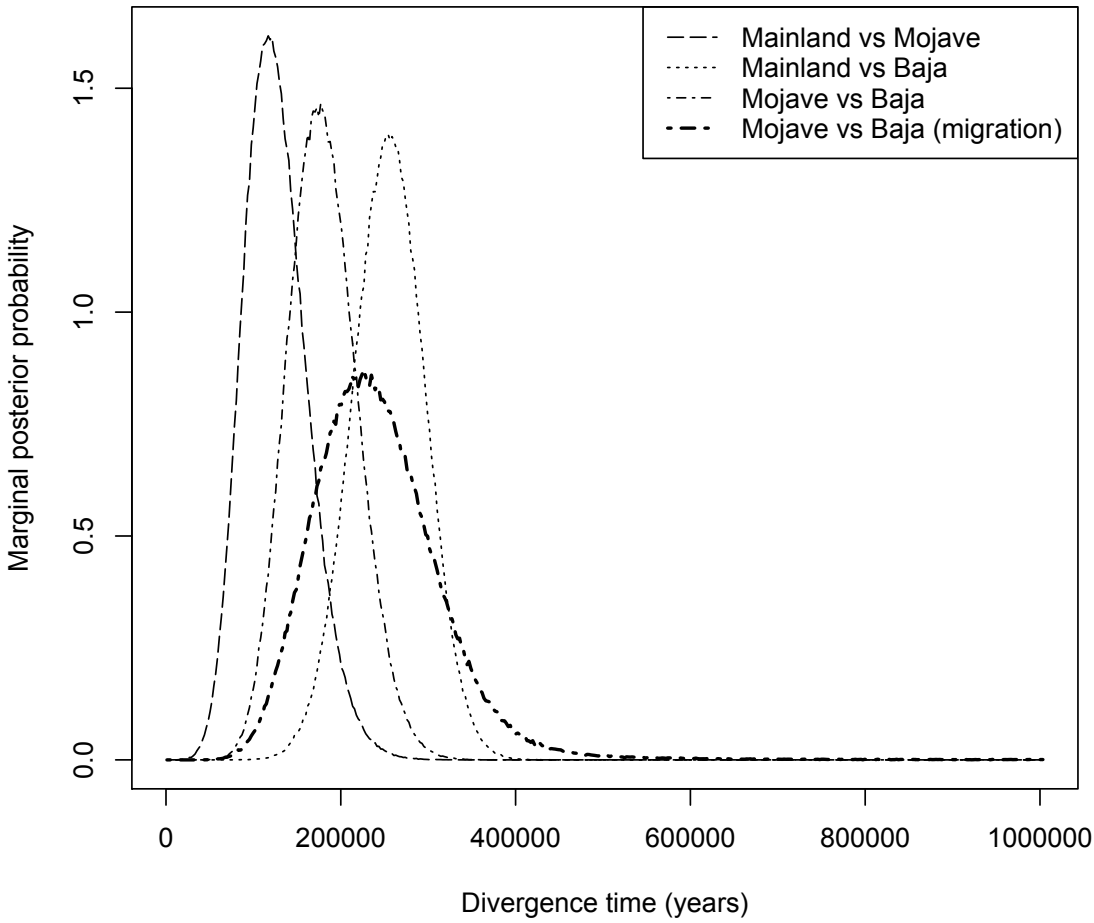


Fig. 3.2 Posterior distributions of divergence time estimates (in years) for each pairwise comparison for the X chromosome dataset. Each line is the posterior distribution for the divergence time parameter under the most parsimonious model (top models in Tables 3.3), for a particular pairwise comparison. The bold distribution is the Mojave Desert/Baja California comparison under Model 2 (equal migration), all others are models without migration.

Consistent with previous studies the effective population size (N_e) of the Mojave Desert was considerably smaller than that of the Mainland population (Table 3.5a). N_e estimates for each of the three populations did not differ between different pairwise comparisons (Fig. 3.3a and Table 3.5a). Similarly, the ancestral N_e of Baja/Mainland was almost identical to that of the Mojave/Baja ancestor (Table 3.5a) regardless of whether migration was included in the model (Fig. 3.3b). This close agreement of

ancestral N_e estimates for both comparisons involving the Baja population is further indication of the basal position of this group in the population history.

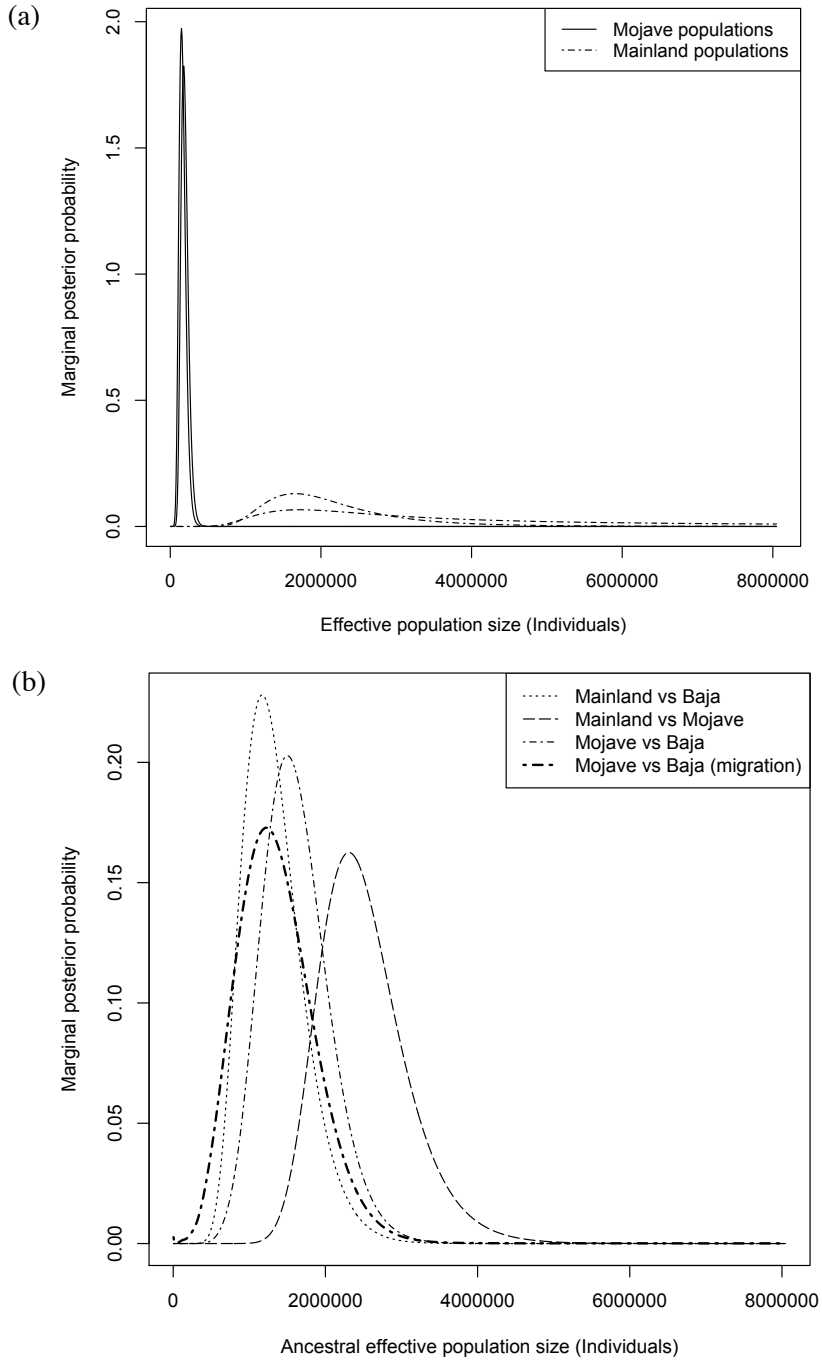


Fig. 3.3 Posterior distributions of effective population size estimates (in number of individuals) for the X chromosome dataset. (a) Contemporary population sizes comprised of two estimates for each population from pairwise comparisons. Baja effective population size (not shown) could not be estimated (see Discussion). (b) Ancestral population sizes for each pairwise comparison, including that of the Model 2 (equal migration) in bold.

Independent of the pairwise comparison or dataset (X-linked or Machado *et al*. 2007) used, the Baja N_e consistently failed to converge in all MCMC runs. We explored this issue by changing the prior for this parameter. However, even at unrealistically high prior bounds posteriors did not converge and we interpret this as a genuine lack of information about this parameter in the data resulting either from a large population size or population sub-structure. Importantly, our explorative runs with varying prior bounds demonstrated that our inability to estimate the Baja N_e did not affect the inference of other model parameters.

Three-population models

The pairwise comparisons in IMa2 suggested a tree topology (Baja, (Mainland, Mojave)) with little gene flow between each population. Using this topology in IMa2, we performed two three-population runs based on the two best models in the Baja/Mojave comparisons; one without migration and one with symmetric migration between the Baja California population and both the Mojave and the ancestor of the Mojave and Mainland populations. With this three-population model migration before and after the Mojave/Mainland split was modelled by separate migration rate parameters providing additional information about the timing of gene flow.

Parameter estimates were almost identical to those of the pairwise comparisons with the addition of migration terms having minor effects on divergence times and N_e estimates. The divergence estimate of the Baja California population from all others was 229 Kya and the Mojave/Mainland split was estimated at 123 Kya (Table 3.5b). Similarly, estimates of N_e were consistent with the pairwise results (Table 3.5b) and ancestral N_e was most sensitive to the inclusion of migration, with lower estimates when migration was included. As before the Baja N_e did not converge.

Estimates of migration were low, with the majority occurring early from the ancestor of Mojave/Mainland to Baja California and very little thereafter (Table 3.5b).

The *BEAST results confirmed the topology and the divergence times estimated by IMa under models without migration (median posterior of 211 Kya and 116 Kya for the two splitting times respectively; Fig. S3.3 and Table 3.5).

Machado et al. (2007) dataset

Model ranking revealed a lack of resolution in the Machado *et al*. (2007) dataset. Both the Baja/Mojave and Baja/Mainland analyses yielded similar evidence ratios for a large number of models with little support for any particular model (Supplementary Tables S3 and S4) and L-mode runs failed for the Mojave/Mainland comparison. However, despite this, the autosomal data were compatible with the inferences from the X data in three key aspects. Firstly, divergence times estimated in all three comparisons broadly agreed (Table 3.5) with the X chromosome results, although posterior distributions had much wider 95% HPD intervals (Fig. S3.1). Secondly, there was little evidence for migration in any comparison and the final simplified model runs were all performed without migration terms. Finally, the Mojave Desert N_e obtained from the Machado *et al*. (2007) data was greatly reduced, although much less so than in the X chromosome data, a pattern not seen for the other two populations (Table 3.5 and Discussion).

DISCUSSION

We analysed sequence data from multiple X chromosome and autosomal loci in a model-based framework to resolve the history of three well-studied, divergent populations of *D. mojavensis*. In particular we inferred both the order of their

divergence and tested for historical gene flow. The X chromosome data suggested that Baja split from an ancestral Mojave/Mainland group 230-270 Kya. The Mojave Desert and Mainland populations diverged more recently, 117-135 Kya, with little subsequent gene flow. Using a three-population model we demonstrated that the weak migration signal detected in the pairwise comparisons was historical, occurring mostly from the common ancestor of the Mojave/Mainland populations to the Baja California population, with a nearly zero rate of migration after the Mojave/Mainland divergence.

A number of studies have examined the phylogeography of *D. mojavensis* using various molecular markers and both population genetic summary statistics and phylogeographic tree based methods (Machado *et al*. 2007; Matzkin 2004; Matzkin & Eanes 2003; Reed *et al*. 2007; Ross & Markow 2006; Ruiz *et al*. 1990; Zouros 1973). Contrary to Ross and Markow (2006) we found the Mojave Desert population to be more closely related to the Mainland but uphold their conclusion that a bottleneck occurred during the history of the Mojave population. While our study found the split between the Baja and Mainland populations to predate that of the Mojave/Mainland, an earlier study based entirely on mtDNA (Reed *et al*. 2007) inferred the opposite order of divergence. Our divergence time estimates coincide with the lower end of the mtDNA estimates, yet the order of divergence differs. This conflicting result illustrates the well-known limitations of single locus data in general, and mtDNA in particular, to resolve recent histories (Ballard & Whitlock 2004) and highlights the superiority of joint inference from multiple loci.

Our estimates of divergence time for the three *D. mojavensis* populations coincide with two interglacial periods, suggesting their history is intimately linked to Pleistocene climatic oscillations. Global temperature oscillations occurred from the

Mid-Pleistocene (900 Kya) onwards, with a periodicity of 100,000 years (Kraaijeveld & Nieboer 2000) and had substantial effects on species formation. In northern hemisphere temperate regions there is a well-characterised pattern of subspecific divergence relating to Pleistocene climate change; however, subtropical regions have been less intensively studied (Hewitt 2011). Climatic as well as vicariant events have influenced the distribution of a range of North American desert species. For example two closely related species pairs of cactophilic *Drosophila*, *D. aldrichi*/*D. wheeleri* and *D. longicornis*/*D. mainlandi*, each show vicariant species distributions caused by the formation of the Sonoran and Mojave Deserts and are separated by the Gulf of California (Beckenbach *et al*. 2008). Some mammal species in these regions have been affected by vicariant events including the formation and movement of the Baja peninsula and the late Pleiocene-Pleistocene uplift of the Mojave Desert region (Bell *et al*. 2010; Riddle *et al*. 2000). Plant species in these regions, including cacti, and insect species have been influenced by both vicariant events and climatic factors (Nason *et al*. 2002; Noonan 1988; Pfeiler & Markow 2011; Thorne 1986). The phylogeographic history of *D. mojavensis* we uncover here is compatible with a previously suggested scenario in which the colonization of mainland Mexico occurred from the Baja peninsula, where *D. mojavensis* originated (Ruiz *et al* 1990; Etges 1999; Matzkin 2004). The isolation of the Baja California population from the ancestor of the Mojave Desert and Mainland populations then followed, approximately 230-270 Kya. This divergence estimate coincides with the global sea level rise of nearly 100m that occurred during the Holstein interglacial (230-215 Kya), which would have separated the Baja peninsula from mainland Mexico (Siddall *et al*. 2007).

We estimate that the Mojave and Mainland populations subsequently subdivided 123-135 Kya during the Eemian interglacial (135-115,000 Kya), after (or during) which the effective size of the Mojave population was reduced by almost an order of magnitude. Previous studies have also inferred a historical bottleneck for the Mojave population (Machado *et al*. 2007; Ross & Markow 2006). Since IMa2 does not explicitly model population bottlenecks, it remains unclear whether this occurred during the divergence from the Mainland population (i.e. through a founder event) or afterwards. The Mojave Desert region is notable for its lack of endemic species with the boundaries of the region defined more by climatic variables rather than geographic features (Axelrod 1983). It is possible that rapid range contractions occurred during the warmer Eemian period led to fragmentation and isolation of the Mojave from the Sonoran population, with a shift to using discontinuous patches of barrel cacti, *F. cylindraceous*, as host plants in this region. Currently, there is some overlap in the distribution of host cacti in these regions, mostly isolated populations of *F. cylindraceous* in southwestern Arizona near populations of organ pipe cacti, *S. thurberi* (Turner *et al*. 1995). However, *D. mojavensis* populations have not been found from east of the Colorado River in southwestern Arizona to the northern range limit of organ pipe cacti in southern Arizona, a gap of ca 200 km (Heed and Etges, unpubl. data). Further, all California and northwestern Arizona populations of *D. mojavensis* are fixed for the second chromosome gene arrangement ST while those in southern Arizona and northwestern Mexico are fixed for LP ($2q^5$), except in one small coastal region in Sonora (Ruiz *et al*. 1990), suggesting no current gene flow between these regions. Our inference of a relatively recent Mainland/Mojave split (123-135 Kya) suggests that there may have been a larger contiguous population in southern

California and Arizona that evolved into discontinuous, differentiated populations that are now fixed for alternate gene arrangements and use different host cacti.

The effective size of the Baja California population appears to be too large to estimate and our analyses consistently failed to produce convergent posteriors for this parameter. Although we cannot say for certain what caused this, it is likely that one or more of the assumptions underlying the IM model was violated. In particular, there is evidence of substructure within regions (Ross & Markow 2006) and AMOVA (Excoffier *et al*. 1992) analyses of population subdivision in Baja California based on inversion karyotype frequencies revealed significant differences among populations grouped by different phytogeographical provinces (Etges *et al*. 1999). While such ecologically driven population structuring may reflect recent/historical conditions, vicariant events have also been implicated in peninsular phylogeographic patterns in several species (Leaché *et al*. 2007). Population structure within the Baja population might be one explanation for an inflated N_e. To test for this, we partitioned the four sampled Baja individuals into northern and southern samples because the Baja peninsula has been geographically divided north to south in the past (Leaché *et al*. 2007). However, including only Northern or Southern Baja into the *BEAST analysis did not change the tree topology or the divergence time estimates. Thus if sub-structure does exist, more intensive within population sampling would be required to reveal it. Alternatively, population size may simply be extremely large. *D. mojavensis* uses its preferred host plant, pitaya agria, almost exclusively on the Baja peninsula. This cactus is known to produce considerably higher densities of breeding sites, "rots", than other host cactus species. In field surveys agria rot densities were ca 40 times higher than those of organ pipe rots (Heed & Mangan 1986), consistent with our estimates of effective population size.

Our results, alongside previous work on the ecological (e.g. Newby and Etges 1998) and genetic (Ruiz *et al* 1990; Etges 1999; Matzkin 2004) relationships of these populations, suggest that the differentiation of the Baja and Mainland involved a host shift due to the colonization of mainland Mexico over the Gulf of California. Taken together, this evidence points towards the Baja California peninsula as being the geographic centre of diversity for this species. Indeed, *D. mojavensis* is likely to have arisen here by allopatric divergence after the ancestor of *D. mojavensis*/*D. arizonae* became separated following the formation of the sea of Cortez, 3-5 Mya (Nason *et al*. 2002; Matzkin 2004). Thus the colonization of the mainland Sonora-Arizona region from the Baja peninsula, during a period of low sea level, would have involved a shift to organ pipe cactus from their preferred agria host. Later, the Baja peninsula and mainland Sonora-Arizona geographic regions were once again isolated when sea levels rose, preventing population gene flow and allowing for divergent selection on life history and reproductive traits on the different cactus hosts (Etges *et al*. 2010; Etges & Heed 1987; Etges *et al*. 1999). Divergence of Baja and Mainland populations in allozyme, inversion and microsatellite frequencies as well as morphology, behaviour, physiology, life history, and host plant use has led some workers to term these geographically isolated groups incipient species, subspecies (Mettler 1963; Pfeiler *et al*. 2009), and races/subraces (Zouros 1973). Further, QTLs for adult epicuticular hydrocarbons (CHCs) which function as contact pheromones in *D. mojavensis*, courtship song and development time are statistically associated with mating success and show genotype-by-environment interactions with host cactus species (Etges *et al*. 2007; Etges *et al*. 2009)**.** Given the timing of subdivision and lack of gene flow we have discovered, it seems that strong host plant associated ecological selection has driven the divergence of life history traits (Etges 1990).

Genetic correlation between egg to adult development time and premating isolation (Etges 1998) and between development time and CHC variation (Etges *et al*. 2010) suggest that adaptation to alternate hosts has caused shifts in different components of male-female courtship signals. Thus, divergent populations of *D. mojavensis* are an important example of how allopatric divergence and ecological selection can indirectly cause divergence in reproductive traits, ultimately leading to reproductive isolation (Funk *et al*. 2006).

Conclusions

Although results obtained from independent analyses on X chromosome and autosomal data were compatible with one another, the X-chromosome data was clearly more informative (Table S3.2). Most likely this was a result of reduced selective constraint on introns (compared to exons) and faster evolution of the X (Schaffner 2004), allowing for greater power to distinguish between models and estimate historical parameters. The estimate of the Mojave Desert N_e obtained from autosomal loci was around twice that of the X chromosome dataset (Table 3.5). Note that we accounted for the difference in N_e between X chromosomes and autosomes (see Methods) and that all other N_e estimates were similar between the two datasets. Such patterns have been observed in the out-of-Africa scenario for both human (Keinan *et al*. 2008) and *D. melanogaster* (Kauer *et al*. 2002) populations, where European populations harbour larger than expected diversity reductions on the X chromosome compared to African populations. While Pool and Nielsen (2007) showed that this could be caused purely by population history (e.g. bottlenecks), other demographic and selective causes cannot be ruled out. Although there is evidence for sex-biased dispersal in some *Drosophila* species (e.g. Begon 1976; Fontdevila and Carson 1978; Powell *et al*. 1976), *D. mojavensis* has not been shown to exhibit this

(Markow and Castrezana 2000). As such we do not expect sex-biased dispersal to have influenced our results.

Although molecular clock calibrations are notoriously prone to errors and uncertainties, they are rarely based on direct measurements of mutation rates nor do they incorporate the uncertainty associated into these estimates as we have done here (Pulquerio & Nichols 2007). While this increases our confidence in the datings obtained in this study, possible uncertainties about generation times and the effect of purifying selection remain. Selective constraint on the evolution of introns is dependent on intron size, with long GC-rich introns potentially having a functional role in gene expression (Haddrill *et al*. 2005). Understanding and incorporating the evolutionary constraints acting on different regions of the genome will much improve phylogeographic inference. While our sampling scheme did resolve the ancestral relationships of the three populations, interesting details of their history, in particular structure within Baja and the bottleneck of the Mojave population, remain to be investigated. However, a more intense sampling scheme, with more individuals, would be required to address these questions.

Inferring the history of closely related populations that may have expanded, contracted, fragmented, or been connected by past migration is fraught with difficulties (Hey & Machado 2003). It has become clear that model-based approaches are required to extract historical signal from DNA sequence data (e.g. Rymer *et al*. 2010). Choosing between the large number of possible models becomes a considerable task for any such analysis when examining more than two populations. We show that a strategy of using model selection on pairwise comparisons to inform a simplified IM model for multiple populations makes this feasible. For phylogeographic histories shaped by Pleistocene climatic events, models of cyclical

rather than continuous gene flow may be more biologically realistic. The recent flood of 'next-generation' sequence data providing entire genomes for analysis (e.g. Forister *et al*. 2010) and increasingly powerful inference methods (e.g. Gutenkunst *et al*. 2010; Lohse *et al*. 2011) will make it possible to further resolve such complex histories.

REFERENCES

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory*, (ed. B.N. P & F C), pp. 267-281, Tsahkadsov, Armenia, USSR.

Anderson DR, Burnham KP (2002) Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management*, **66**, 912-918.

Axelrod DI (1983) *Paleobotanical history of the western deserts* University of New Mexico Press, Albuquerque.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, **13**, 729-744.

Beckenbach AT, Heed WB, Etges WJ (2008) A mitochondrial DNA analysis of vicariant speciation in two lineages in the *Drosophila mulleri* subgroup. *Evolutionary Ecology Research*, **10**, 475-492.

Begon M (1976) Dispersal, density and microdistribution in *Drosophila subobscura* Collin. *Journal Of Animal Ecology,* **45**, 441-456.

Bell KC, Hafner DJ, Leitner P, Matocq MD (2010) Phylogeography of the ground squirrel subgenus *Xerospermophilus* and assembly of the Mojave Desert biota. *Journal of Biogeography*, **37**, 363-378.

Brazner JC, Etges WJ (1993) Pre-mating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. II. Effects of larval substrates on time to copulation, mate choice and mating propensity. *Evolutionary Ecology*, **7**, 605-624.

Carstens BC, Stoute HN, Reid NM (2009) An information theoretical approach to phylogeography. *Molecular Ecology*, **18**, 4270-4282.

Drummond A, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Etges WJ (1990) Direction of life history evolution in *Drosophila mojavensis*. In: *Ecological and Evolutionary Genetics of Drosophila*, (eds. Barker JSF, Starmer WT, MacIntyre RJ), pp. 37-56. Plenum, New York.

Etges WJ (1992) Premating isolation is determined by larval substrates in cactophilic *Drosophila mojavensis*. *Evolution*, **46**, 1945-1950.

Etges WJ (1998) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. IV. Correlated responses in behavioral isolation to artificial selection on a life-history trait. *American Naturalist*, **152**, 129-144.

Etges WJ, de Oliveira CC, Gragg E, *et al*. (2007) Genetics of incipient speciation in *Drosophila mojavensis*. I. Male courtship song, mating success, and genotype x environment interactions. *Evolution*, **61**, 1106-1119.

Etges WJ, de Oliveira CC, Noor MAF, Ritchie MG (2010) Genetics of incipient speciation in *Drosophila mojavensis*. III. Life history divergence in allopatry and reproductive isolation. *Evolution*, **64**, 3549-3569.

Etges WJ, de Oliveira CC, Ritchie MG, Noor MAF (2009) Genetics of incipient speciation in *Drosophila mojavensis*. II. Host plants and mating status influence cuticular hydrocarbon QTL expression and G× E interactions. *Evolution*, **63**, 1712-1730.

Etges WJ, Heed WB (1987) Sensitivity to larval density in populations of *Drosophila mojavensis* - influences of host plant variation on components of fitness. *Oecologia*, **71**, 375-381.

Etges WJ, Johnson WR, Duncan GA, Huckins G, Heed WB (1999) Ecological genetics of cactophilic *Drosophila*. In: *Ecology of Sonoran Desert Plants and Plant Communities*. (ed. Robichaux R), pp. 164-214. University of Arizona Press, Tucson.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotyes: Application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479-491.

Farine JP, Legal L, Moreteau B, Le Quere JL (1996) Volatile components of ripe fruits of *Morinda citrifolia* and their effects on *Drosophila*. *Phytochemistry*, **41**, 433-438.

Fellows DP, Heed WB (1972) Factors affecting host plant selection in desert-adapted cactophilic *Drosophila*. *Ecology*, **53**, 850-858.

Fontdevila A, Carson HL (1978) Spatial distribution and dispersal in a population of *Drosophila*. *The American Naturalist,* **112**, 365-380.

Forister ML, Gompert Z, Fordyce JA, Nice CC (2010) After 60 years, an answer to the question: what is the Karner blue butterfly? *Biology Letters*, **7**, 399-402.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693-709.

Funk DJ, Nosil P, Etges WJ (2006) Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3209-3213.

Garrick RC, Caccone A, Sunnucks P (2010) Inference of population history by coupling exploratory and model-driven phylogeographic analyses. *International Journal of Molecular Sciences*, **11**, 1190-1227.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2010) Diffusion Approximations for Demographic Inference: DaDi.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P (2005) Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biology*, **6**, R67.

Heed WB (1978) Ecology and genetics of Sonoran Desert *Drosophila*. In: *Ecological genetics: The interface*. (ed. Brussard PF), pp. 109-126. Springer-Verlag, New York.

Heed WB, Mangan RL (1986) Community ecology of the Sonoran Desert *Drosophila*. In: *The Genetics and Biology of Drosophila* (eds. Ashburner M, Carson HL, Thompson JN), pp. 311-345. Academic Press, New York.

Hewitt GM (2011) Quaternary phylogeography: the roots of hybrid zones. *Genetica*, **139**, 1-22.

Hey J (2005) On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology*, **3**, e193.

Hey J (2010) Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, **27**, 905-920.

Hey J, Machado CA (2003) The study of structured populations - new hope for a difficult and divided science. *Nature Reviews Genetics*, **4**, 535-543.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747.

Jaenike J (1990) Host specialization in phytophagous insects. *Annual Review of Ecology and Systematics*, **21**, 243-273.

Johnson WR (1980) *Chromosomal polymorphism in natural populations of the desert adapted species Drosophila mojavensis*. PhD thesis, University of Arizona.

Kambysellis MP, Ho KF, Craddock EM, *et al*. (1995) Pattern of ecological shifts in the diversification of Hawaiian *Drosophila* inferred from a molecular phylogeny. *Current Biology*, **5**, 1129-1139.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059-3066.

Kauer M, Zangerl B, Dieringer D, Schlötterer C (2002) Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. *Genetics*, **160**, 247-256.

Keightley PD, Trivedi U, Thomson M, *et al*. (2009) Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research*, **19**, 1195-1201.

Keinan A, Mullikin JC, Patterson N, Reich D (2008) Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nature Genetics*, **41**, 66-70.

Kraaijeveld K, Nieboer EN (2000) Late Quaternary paleogeography and evolution of arctic breeding waders. *Ardea*, **88**, 193-205.

Leaché AD, Crews SC, Hickerson MJ (2007) Two waves of diversification in mammals and reptiles of Baja California revealed by hierarchical Bayesian analysis. *Biology Letters*, **3**, 646-650.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451-1452.

Lohse K, Harrison RJ, Barton NH (2011) A general method for calculating likelihoods under the coalescent process. *Genetics* **189**, 977-987.

Lohse K, Sharanowski B, Stone GN (2010) Quantifying the Pleistocene history of the oak gall parasitoid *Cecidostiba fungosa* using twenty intron loci. *Evolution*, **64**, 2664-2681.

Maan ME, Seehausen O (2011) Ecology, sexual selection and speciation. *Ecology Letters*, **14**, 591-602.

Machado CA, Matzkin LM, Reed LK, Markow TA (2007) Multilocus nuclear sequences reveal intra and interspecific relationships among chromosomally polymorphic species of cactophilic *Drosophila*. *Molecular Ecology*, **16**, 3009-3024.

Magnacca KN, Foote D, O'Grady PM (2008) A review of the endemic Hawaiian *Drosophilidae* and their host plants. *Zootaxa*, **1728**, 1-58.

Mallet J (2006) What does *Drosophila* genetics tell us about speciation? *Trends in Ecology & Evolution*, **21**, 386-393.

Markow TA (1991) Sexual isolation among populations of *Drosophila mojavensis*. *Evolution*, **45**, 1525-1529.

Markow TA, Castrezana S (2000). Dispersal in cactophilic *Drosophila*. *Oikos,* **89**, 378-386.

Matzkin LM (2004) Population genetics and geographic variation of alcohol dehydrogenase (*Adh*) paralogs and glucose-6-phosphate dehydrogenase (*G6pd*) in *Drosophila mojavensis*. *Molecular Biology and Evolution*, **21**, 276-285.

Matzkin LM, Eanes WF (2003) Sequence variation of alcohol dehydrogenase (*Adh*) paralogs in cactophilic *Drosophila*. *Genetics*, **163**, 181-194.

Mettler LE (1963) *Drosophila mojavensis baja*, a new form in Mulleri complex. *Drosophila Information Service*, **38**, 421-430.

Newby BD, Etges WJ (1998) Host preference among populations of *Drosophila mojavensis* (Diptera: Drosophilidae) that use different host cacti. *Journal of insect behaviour*. **11**, 691-712.

Nason JD, Hamrick J, Fleming TH (2002) Historical vicariance and postglacial colonization effects on the evolution of genetic structure in *Lophocereus,* a Sonoran Desert columnar cactus. *Evolution*, **56**, 2214-2226.

Nei M (1987) *Molecular evolutionary genetics* Columbia University Press, New York.

Noonan GR (1988) Biogeography of North American and Mexican insects, and a critique of vicariance biogeography. *Systematic Zoology*, **37**, 366-384.

Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Molecular Biology and Evolution*, **22**, 2119-2130.

Pfeiler E, Castrezana S, Reed L, Markow T (2009) Genetic, ecological and morphological differences among populations of the cactophilic *Drosophila mojavensis* from southwestern USA and northwestern Mexico, with descriptions of two new subspecies. *Journal Of Natural History*, **43**, 923-938.

Pfeiler E, Markow TA (2011) Phylogeography of the cactophilic *Drosophila* and other arthropods associated with cactus necroses in the Sonoran Desert. *Insects*, **2**, 218-231.

Piganeau G, Eyre-Walker A (2004) A reanalysis of the indirect evidence for recombination in human mitochondrial DNA. *Heredity*, **92**, 282-288.

Piganeau G, Gardner M, Eyre-Walker A (2004) A broad survey of recombination in animal mitochondria. *Molecular Biology and Evolution*, **21**, 2319-2325.

Pool JE, Nielsen R (2007) Population size changes reshape genomic patterns of diversity. *Evolution*, **61**, 3001-3006.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253-1256.

Powell JR, Dobzhansky T, Hook JE & Wistrand HE (1976) Genetics of natural populations. XLIII. Further studies on rates of dispersal of *Drosophila pseudoobscura* and its relatives. *Genetics,* **82**, 493.

Pulquerio MJF, Nichols RA (2007) Dates from the molecular clock: how wrong can we be? *Trends in Ecology & Evolution*, **22**, 180-184.

R'kha S, Capy P, David JR (1991) Host-plant specialization in the *Drosophila melanogaster* species complex: a physiological, behavioral, and genetical analysis. *Proceedings of the National Academy of Sciences*, **88**, 1835.

Rambaut A, (2010) FigTree 1.3.1. [http://tree.bio.ed.ac.uk/software/figtree/].

Reed LK, Nyboer M, Markow TA (2007) Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. *Molecular Ecology*, **16**, 1007-1022.

Riddle BR, Hafner DJ, Alexander LF, Jaeger JR (2000) Cryptic vicariance in the historical assembly of a Baja California Peninsular Desert biota. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 14438-14443.

Ross C, Markow T (2006) Microsatellite variation among diverging populations of *Drosophila mojavensis*. *Journal of Evolutionary Biology*, **19**, 1691-1700.

Ruiz A, Heed W, Wasserman M (1990) Evolution of the *mojavensis* cluster of cactophilic *Drosophila* with descriptions of two new species. *Journal of Heredity*, **81**, 30-42.

Ruiz A, Heed WB (1988) Host-plant specificity in the cactophilic *Drosophila mulleri* species complex. *The Journal of Animal Ecology*, **57**, 237-249.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336-352.

Rymer PD, Manning JC, Goldblatt P, Powell MP, Savolainen V (2010) Evidence of recent and continuous speciation in a biodiversity hotspot: a population genetic approach in southern African gladioli (Gladiolus; Iridaceae). *Molecular Ecology*, **19**, 4765-4782.

Schluter D (2001) Ecology and the origin of species. *Trends in Ecology & Evolution*, **16**, 372-380.

Schaffner SF (2004) The X chromosome in population genetics. *Nature Reviews Genetics*, **5**, 43-51.

Siddall M, Chappell J, Potter EK (2007) Eustatic sea level during past interglacials. In: *The climate of past interglacials*. (eds. Sirocko F, Claussen M, Sanchez Goni MF, Litt T), pp. 75-92. Elsevier, London.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, **141**, 413.

Sobel JM, Chen GF, Watt LR, Schemske DW (2010) The biology of speciation. *Evolution*, **64**, 295-315.

Stennett MD, Etges WJ (1997) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. III. Epicuticular hydrocarbon variation is determined by use of different host plants in *Drosophila*

mojavensis and *Drosophila arizonae*. *Journal of Chemical Ecology*, **23**, 2803-2824.

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

Thorne RF (1986) A historical sketch of the vegetation of the Mojave and Colorado deserts of the American Southwest. *Annals Of the Missouri Botanical Garden*, **73**, 642-651.

Turner RM, Bowers JE, Burgess TL (1995) Sonoran Desert plants: an ecological atlas. *University of Arizona Press, Tucson*.

Van Doorn GS, Edelaar P, Weissing FJ (2009) On the origin of species by natural and sexual selection. *Science*, **326**, 1704-1707.

Wang Y, Hey J (2010) Estimating divergence parameters with small samples from a large number of loci. *Genetics*, **184**, 363-379.

Wasserman M (1992) Cytological evolution of the *Drosophila repleta* species group. In: *Inversion polymorphism in Drosophila*. (eds. Powell JR, Krimbas CB), pp. 455-541. CRC Press Inc., Bota Racon, Florida.

Watterson G (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256-276.

Weissing FJ, Edelaar P, van Doorn GS (2011) Adaptive speciation theory: a conceptual review. *Behavioral Ecology and Sociobiology*, 1-20.

Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: Practice, problems and prospects. *Molecular Ecology*, **12**, 563-584.

Zouros E (1973) Genic differentiation associated with the early stages of speciation in the mulleri subgroup of *Drosophila*. *Evolution*, **27**, 601-621.

Zouros E, d'Entremont CJ (1980) Sexual isolation among populations of *Drosophila mojavensis*: response to pressure from a related species. *Evolution*, **34**, 421-430.

CHAPTER 4:

TRANSCRIPTOME-WIDE EXPRESSION VARIATION ASSOCIATED WITH ENVIRONMENTAL PLASTICITY AND MATING SUCCESS IN CACTOPHILIC *DROSOPHILA MOJAVENSIS*

MATERIAL FROM THIS CHAPTER PUBLISHED AS:

Smith G, Fang Y, Liu X, Kenny J, Cossins AR, de Oliveira CC, Etges WJ & Ritchie MG (2013) Transcriptome-wide expression variation associated with environmental plasticity and mating success in cactophilic *Drosophila mojavensis*. *Evolution*. (*Online first*).

AUTHOR CONTRIBUTIONS:

GS and MGR conceived the project. WJE and CCO provided flies and performed rearing and mating success experiments. GS performed RNA extraction and preparation. JK performed RNA sequencing, XL and YF performed bioinformatic and statistical analyses with GS. GS wrote the paper with contributions from all co-authors.

ABSTRACT

Ecological speciation occurs with the adaptation of populations to different environments and concurrent evolution of reproductive isolation. Phenotypic plasticity might influence both ecological adaptation and reproductive traits. We examined environment-specific gene expression and male mating success in cactophilic *Drosophila mojavensis* using transcriptome sequencing. This species exhibits cactus-dependent mating success across different species of host plants, with genotype-by-environment interactions for numerous traits. We cultured flies from egg to eclosion on two natural cactus hosts and surveyed gene expression in adult males that were either successful or unsuccessful in achieving copulation in courtship trials. We identified gene expression differences that included functions involved with metabolism, most likely related to chemical differences between host cactus species. Several epigenetic-related functions were identified that might play a role in modulating gene expression in adults due to host cactus effects on larvae, and mating success. Cactus-dependent mating success involved expression differences of genes implicated in translation, transcription and nervous system development. This suggests a role of neurological function genes in the mating success of *D. mojavensis* males. Together, these results suggest that the influence of environmental variation on mating success *via* regulation of gene expression might be an important aspect of ecological speciation.

INTRODUCTION

Environmentally induced phenotypic variation is an under-studied aspect of speciation, particularly its role in the evolution of reproductive isolation (Butlin *et al*. 2012). However, phenotypic plasticity provides a way for the same genotype to produce different traits according to environmental context and thus may be an important process facilitating ecological adaptation and speciation (Thibert-Plante & Hendry 2011). Assessment of the role of phenotypic plasticity in ecological adaptation is difficult due to a poor understanding of the developmental processes producing plastic traits and how these link to environmental adaptation (Ghalambor *et al*. 2007; Scoville & Pfrender 2010). If an organism possesses the developmental flexibility necessary for a plastic response then ecological adaptation may result in particularly rapid divergence due to ecological selection on relevant traits (Schluter 2001; Rundle & Nosil 2005). When such traits diverge, speciation may follow if reproductive isolation evolves as a direct or indirect consequence (Funk *et al*. 2006).

Speciation may be more likely when ecological and sexual selection interact (Ritchie 2007; Van Doorn *et al*. 2009; Sobel *et al*. 2010; Maan & Seehausen 2011), if traits that influence mating success are under divergent ecological selection. Determining how intimate any such associations are (for example, if this involves pleiotropy or covariance between traits) is central to understanding debates about the likelihood of so-called 'magic trait' speciation (Servedio *et al*. 2011) or how extensive traits with multiple effects are (Smadja & Butlin 2011). Plasticity of ecological and reproductive traits may be important for adaptation and speciation if selection in an environment facilitates genetic population divergence, perhaps *via* genetic accommodation (West-Eberhard 2005a; West-Eberhard 2005b), or adaptive plasticity could simply allow persistence for long enough to allow genetic adaptation (Crispo

2008). There are multiple ways in which plasticity could either accelerate or inhibit speciation (Pfennig *et al*. 2010) but most current methodologies to detect genes involved in ecological adaptation ignore expression variation (Pavey *et al*. 2010).

Drosophila species are commonly utilized for both gene expression and plasticity studies (Levine *et al*. 2011). The cactophilic species, *Drosophila mojavensis*, is an excellent candidate for such studies due to the use of different host cacti during development, and has been used as a model system for divergence during speciation. This species is endemic to northwestern Mexico and the southwestern United States and its range is comprised of four allopatric regions with little evidence of contemporary gene flow between them (Ross & Markow 2006; Machado *et al*. 2007; Reed *et al*. 2007). Populations of *D. mojavensis* that occupy Baja California and the mainland Mexico-Arizona regions are of particular interest as they demonstrate a significant level of environmentally influenced premating sexual isolation between them, mediated by divergent epicuticular hydrocarbon (CHC) profiles, which act as contact pheromones, and courtship songs (Etges *et al*. 2007; Etges *et al*. 2009). This premating isolation is seen only between the Baja and mainland populations, with little evidence of postzygotic isolation between any population pair (Ruiz *et al*. 1990). The Baja California peninsula is thought to be where *D. mojavensis* originated, based on genetic and ecological evidence (Ruiz *et al*. 1990; Wasserman 1992; Matzkin & Eanes 2003; Matzkin 2004), with all populations diverging from an ancestral group around 230-270,000 years ago (Smith *et al*. 2012). Colonization of the mainland Mexico involved a host plant shift from the favored pitaya agria cactus, *Stenocereus gummosus*, to organ pipe cactus, *S. thurberi*, that is distributed in the southern half of Baja California, mainland Sonora and Sinaloa and southern Arizona (Heed 1982; Etges *et al*. 1999). A suite of phenotypic changes in life history and reproductive traits

accompanied this host shift and many of these phenotypes demonstrate plasticity when flies are raised on differing cactus hosts (Etges *et al*. 2007; Etges *et al*. 2009; Etges *et al*. 2010). Therefore *D. mojavensis* has undergone ecological selection with concurrent divergent selection of reproductive traits, and shows plastic expression of many key traits across cactus hosts.

Plasticity due to gene expression variation is potentially an important component of adaptation to a varying environment and can act alongside gene-environment interactions in determining levels of adaptation during ecological specialisation. A series of Quantitative Trait Locus (QTL) studies revealed the genetic architecture of life history traits involved in host plant adaptation and sexual isolation in *D. mojavensis* (Etges *et al*. 2007; Etges *et al*. 2009; Etges *et al*. 2010). F_2 males from crosses between Baja and mainland populations were reared on either agria or organ pipe cactus and QTLs identified for mating success, courtship songs, cuticular hydrocarbons and egg to adult development time. All traits showed evidence of genotype-by-environment interactions (GxEs) influencing their expression. Hence this system displays extensive genetic variability influencing viability and mating success. The extent of variation in gene expression due to rearing cactus (or other environmental variation) in *D. mojavensis* is only beginning to be studied. Microarray analyses revealed approximately 1500-3000 genes with cactus-specific expression in third instar larvae in Baja and Mainland populations reared on agria vs. organ pipe cacti (Matzkin *et al*. 2006; Matzkin 2012), thousands of genes in adults under dessication stress (Rajpurohit *et al*. 2013, *in press*) and thousands of genes assayed across the entire life cycle (Etges WJ, unpubl. data). Many of these genes were involved in metabolism and detoxification pathways as would be predicted due to

host plant chemical differences (Fogleman and Danielson 2001), as well as fatty acid biosynthesis and olfaction (Matzkin *et al*. 2006; Matzkin 2012).

Here we describe an RNA sequencing study of gene expression variation associated with different cactus substrates in adult *D. mojavensis* that explicitly examines the link between ecological variation and courtship behaviour. We examined gene expression in adult males from a mainland population reared from egg to eclosion on either organ pipe or agria cactus, after identifying the first males to succeed in mating trials with mainland females. Our aim was to 1) enumerate and identify genes or functional gene networks that showed plastic expression responses to host cactus, 2) identify expression variation associated with rapid mating success, and 3) examine the interaction between cactus and mating success variation. The latter is particularly important for identifying genes and functional pathways involved in cactus-dependent mating success; for example pleiotropic linkage between cactus adaptation and mating behavior would predict that the same genes are involved in both traits. Note that we reared flies to fermenting cactus from egg to eclosion and surveyed expression differences in adults; here we use the term 'epigenetics' to include induced gene expression changes during development (e.g. Chittka *et al*. 2012) and when used does not imply that we have identified trans-generational effects.

METHODS

Fly maintenance

All experiments were performed with a population of *D. mojavensis* from Organ Pipe National Monument (OPNM), Arizona, collected in 2002 by T. Markow. This multifemale stock was reared en mass on banana food in 8 dr shell vials at ambient temperature. Although this population was known to be homokaryotypic for gene

arrangements on the second and third chromosomes (Etges *et al*. 1999), we made multiple pair-mated lines and cytologically verified that no inversions were segregating. We then sib-mated these lines for five generations and one inbred line was selected for the mating trials described below. Flies from this inbred line were derived from an isofemale line established in 2004. This is the same line used in the QTL crosses analyzed recently (Etges *et al*. 2007; Etges *et al*. 2009; Etges *et al*. 2010). We chose to analyze the mainland line because this derived population has successfully performed a host shift, and females from the mainland are more discriminating in mate choice (Markow 1991; Etges 1992). Flies were reared on banana food at moderate larval densities in half-pint bottles in an incubator at 27 °C during the day and 17 °C at night on a 14-h light:10-h dark cycle. Emerging adults were aged until sexually mature (10-12 days) then placed into oviposition chambers (~400 adults per chamber), allowed to mate and then oviposit for 10 h each day. Eggs were washed in deionized water, 70% ethanol and again in sterile deionized water. Groups of 200 eggs were transferred to a 1 cm^2 piece of sterilized filter paper and then placed on fermenting cactus tissue, either agria, *S. gummosus*, or organ pipe cactus, *S. thurberi*, in an incubator programmed as above. Experimental flies were reared on each cactus species from egg to eclosion, and thereafter on banana food until sexual maturity.

Fermenting cactus cultures were set up in half pint bottles with 75 g of aquarium gravel covered with a 5.5-cm diameter piece of filter paper. Bottles were autoclaved, 60 g of either agria or organ pipe tissues were added then autoclaved again for 8 min at low pressure (Etges 1998). After cooling to room temperature, each culture was inoculated with 0.5 mL of a pectolytic bacterium, *Erwinia cacticida* (Alcorn *et al*. 1991) and 1.0 mL of a mixture of seven yeast species common in

natural agria and organ pipe rots (Starmer 1982): *Dipodascus starmeri*, *Candida sonorensis*, *C. valida*, *Starmera amethionina*, *Pichia cactophila*, *P. mexicana*, and *Sporopachydermia cereana*. Inoculation was performed to ensure that microorganisms found in nature, rather than those present in the laboratory environment, populated the necrotic tissue, and differential growth of the inoculants will generate variation between cactus types. Eggs were then added and, once eclosed, adults from four replicate cactus bottles were separated by gender and kept on banana food in vials in the incubator until sexually mature at 10-12 days.

Mate choice experiments

We identified males who were successful in mating trials to assess the influence of preadult cactus rearing on the mating behaviour of male flies. Males used in the mate trials were reared on agria or organ pipe cactus, whereas all females were reared only on organ pipe cactus, in order to identify the effects of rearing substrates on male mating success. Organ pipe cactus reared females are more discriminating in mate choice trials than agria cactus-reared females (Etges 1992). A total of 4 treatments (two rearing cacti and two mating statuses) were performed, with four replicates per treatment.

Mate choice trials were carried out using a multiple-choice design (Etges 1992; Etges & Ahrens 2001). Twenty female and male virgin adults 12 to 16 days old were used in each trial. A 50 mL Erlenmeyer flask was used as a mating chamber and changed after each trial. Each trial lasted until half of the pairs copulated or for a maximum of half an hour. All trials were performed at room temperature (18 - 20° C) in the morning (10:00 AM to 12:00 PM; lights on at 6AM) over no more than a two-hour time period. We define the first mating males as 'successful' and the non-mating

as non-successful, so our measure of mating success potentially includes measures of male mating speed and vigour as well as female discrimination. However, previous studies suggest that male mating success is due overwhelmingly to female choice and not male-male interaction or unwillingness to mate (Havens *et al*. 2011). In mating trials with cactus-reared adults, usually almost all flies mate by the end of the trials. Copulating pairs were observed for at least 10 seconds to avoid any pseudo-copulating pairs (Markow *et al*. 1983), then flash frozen in liquid nitrogen and stored in RNAlater at -20° C prior to RNA extraction. After half of the flies had mated, the remaining unmated flies were also flash frozen. Males from two mate choice trials, i.e. 20 male whole bodies, were pooled together per sample.

RNA sequencing and analysis

RNA was extracted from each pooled sample using a Qiagen RNeasy Mini Kit (Qiagen, Valencia, California USA), and the high quality of the RNA was confirmed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, California). Library preparation was carried out at the Centre for Genomic Research, University of Liverpool and included poly(A) tail selection using the Dynabeads mRNA purification kit (Invitrogen, Paisley, UK). 100ng of the resulting mRNA was used as input for library production using the SOLiD Total RNA-Seq kit (Life Technologies, Paisley, UK). Sequencing was performed on a SOLiD 4 sequencer (Life Technologies), generating reads of 50bp length. The sequence data has been deposited in the European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EBI, accession ERP002218).

Reads were filtered for quality and mapped on to the *D. mojavensis* genome (Clark et al. 2007) using the Tuxedo suite (Bowtie/Tophat v1.4.0), with only uniquely

mapped reads retained for analysis. HTSeq (Anders 2010) was used to quantify read counts for two different types of feature; genes and exons. This enabled a quantitative examination of expression levels at the level of whole genes as well single exons within a gene.

Raw count data were normalized across sequenced libraries and a generalized linear model was fitted with a negative binomial distribution using the EdgeR package in R (Robinson et al. 2010). The full model was $Y = g^{-1}(AX) + \varepsilon$ where Y is a vector of normalized count numbers for one gene, $g^{-1}(.)$ is the inverse link function, A is the model matrix and X incorporates the model parameters. X includes an intercept and effects of factors; cactus (organ pipe or agria), mating success (success or fail) and an interaction effect, whilst ε represents the random noise of sampling a negative binomial population. A likelihood ratio test was used to compare a full model to models with each term removed, to test for significance of each treatment effect. Gene specific *p*-values were corrected for multiple testing using the False Discovery Rate (FDR) approach of Storey & Tibshirani (2003) with significance taken at 10%. A heatmap was constructed for significantly differentially expressed (DE) genes overall and clusters of co-regulated genes were obtained using the k-means clustering (Hartigan & Wong 1979) from the R package.

Analysis at the level of exons can identify those genes where alternative exon expression arising from alternative transcript initiation, alternative splicing or alternative polyadenylation occurs due to treatment (Griffith *et al*. 2010). For this we used a modified version of the function *spliceVariants* from the EdgeR package. This fits a negative binomial generalized linear model for each gene, given the counts for the exons within that gene. The same approach was adopted as for the gene-level testing except that term X (see previous paragraph) now includes terms reflecting the

difference among the exons, treatment effects and interaction of exon and treatment. Thus the model becomes $E(Z)=g^{-1}(B\beta)+\varepsilon$, where ε is random noise, Z is a vector of normalized count numbers for all exons within one gene, $g^{-1}(.)$ is the same as stated before, B is the model matrix, β is the vector of model parameters which includes the intercept and terms reflecting the difference among the exons, treatment effects and interaction of exon and treatment. Because exon expression is modeled according to the other exons within the same gene, significance represents independent responses of exons by treatment, or alternative expression (AE). That is, the test reveals genes for which there is an exon-specific signature across treatments or interactions between exons and treatments. Significant AE was taken for each gene at 10% FDR.

Functional enrichment

Functional enrichment or over-representation analysis aims to detect common functions within the DE or AE gene sets. For categorizing genes we used Gene Ontology (GO) annotations from the *D. mojavensis* entries in QuickGO (Binns *et al*. 2009), as well using corresponding GO terms for orthologous genes in *D. melanogaster* FlyBase entries (vFB2012_01, *D. melanogaster* release 5.43) taken from the ortholog conversion tool (McQuilton *et al*. 2012).

Two test methods were employed, and each were undertaken separately for both DE and AE genes across the three contrasts (cactus, mating success and their interaction) using FDR < 10%. First, we developed a 'rank mean test' written in R (*marrayRankTest*, available from Y. Fang), which ranks genes by *p*-value and takes the mean of the rank for members of gene sets (i.e. biological-process GO categories, downloaded from QuickGO; Binns *et al*. 2009) as the test statistic for enrichment. This method employs a corrected normal distribution that more accurately estimates

p-values for enrichment with small gene sets, which can occur with poorly annotated genomes.

Second, *D. melanogaster* orthologs of the responding DE and AE genes (FDR < 10%) were analysed in DAVID v6.7 (Database for Annotation, Visualization and Integrated Discovery v6.7; Dennis *et al*. 2003; Huang *et al*. 2009). DAVID uses Fisher's exact test to identify significantly enriched GO categories. A 'Fuzzy' clustering algorithm then groups annotation terms into functional clusters of genes. Clusters are considered significant with an Enrichment Score (the geometric mean of annotation *p*-values) >1.3 (Dennis *et al*. 2003; Huang *et al*. 2009).

Several small, non-coding RNAs (ncRNAs) were identified as significantly DE. For each, the orthologous ncRNA in *D. melanogaster* was found through a BLASTn search on the NCBI website, using the nucleotide database (Altschul *et al*. 1997). Hits were called as being significantly orthologous when the E-value was <10e-6, and GO annotations for each ncRNA were obtained from FlyBase.

RESULTS

The experimental design consisted of a pooled sample of 20 males for each of the 2 factors (mating success) and 2 treatment (rearing cactus diet) groups. Four biological replicates were produced independently from rearing to mating success producing 16 RNA samples each of which was subjected to RNA-Seq. Fragment data was analysed by ANOVA for differential gene expression (DE) and alternative expression (AE), each generating three contrasts between rearing cacti, differential mating success and their interaction. The number of reads obtained per sample was typically 45 million reads of 50bp length. Of these, approximately 30% mapped uniquely to the reference genome, producing a final average of 15.3 million reads per biological replicate. The

number of DE genes was relatively small with 111 showing main effects of cactus, 19 associated with mating success and 147 due to their interaction (Table 4.1), all from 212 unique gene models. Fewer genes were generated by the AE analysis, with 64 and 48 for cactus and interaction effects, respectively. However in contrast to DE genes, there was a greater involvement of alternative expression in male mating success (28; Table 4.1).

Table 4.1 Number of genes with significant cactus, mating success and interaction effects. The number of differentially expressed (DE) and alternatively expressed (AE) genes in *D. mojavensis* are presented, along with the number of genes that had confirmed *D. melanogaster* orthologs and the total number of unique genes across all effects.

Test	Cactus	Mating Success	Interaction	Total unique gene models
DE *D. mojavensis*	111	19	147	212
DE *D. melanogaster*	71	11	92	144
AE *D. mojavensis*	64	28	48	115
AE *D. melanogaster*	51	24	38	100

Differential expression

Fig. 4.1 indicates the fold-change responses of all DE genes as a heatmap across the three ANOVA contrasts. K means clustering generated 8 clusters; clusters 3-8 were up-regulated by the organ pipe relative to agria diets, and clusters 5, 6, 3 and 8 were up-regulated and 1, 2 and 7 were down-regulated by successful relative to unsuccessful mating. Full details of all DE and AE genes and all functional enrichment results are given in supplementary Tables S4.1-S4.6, and GO term annotations for genes within each heatmap cluster in Table S4.7.

Functional enrichment of cactus-specific DE genes showed functions for several processes linked to metabolism (Fig. 4.2). These included glycerol ether metabolic processes, the tricarboxylic acid (TCA) cycle and cell redox homeostasis. Annotations for protein modification were also significant, specifically protein

ubiquitination and ubiquitin mediated protein catabolism. Other significant annotations included immune response, methylation, tRNA processing and calcium-based signaling. Genes showing expression differences by male mating success were enriched for two terms; translation and glycerol ether metabolic process (Fig. 4.2). The interaction of cactus and mating success produced only one significant annotation, for translation (Fig. 4.2), despite having the greatest number of significantly DE genes.

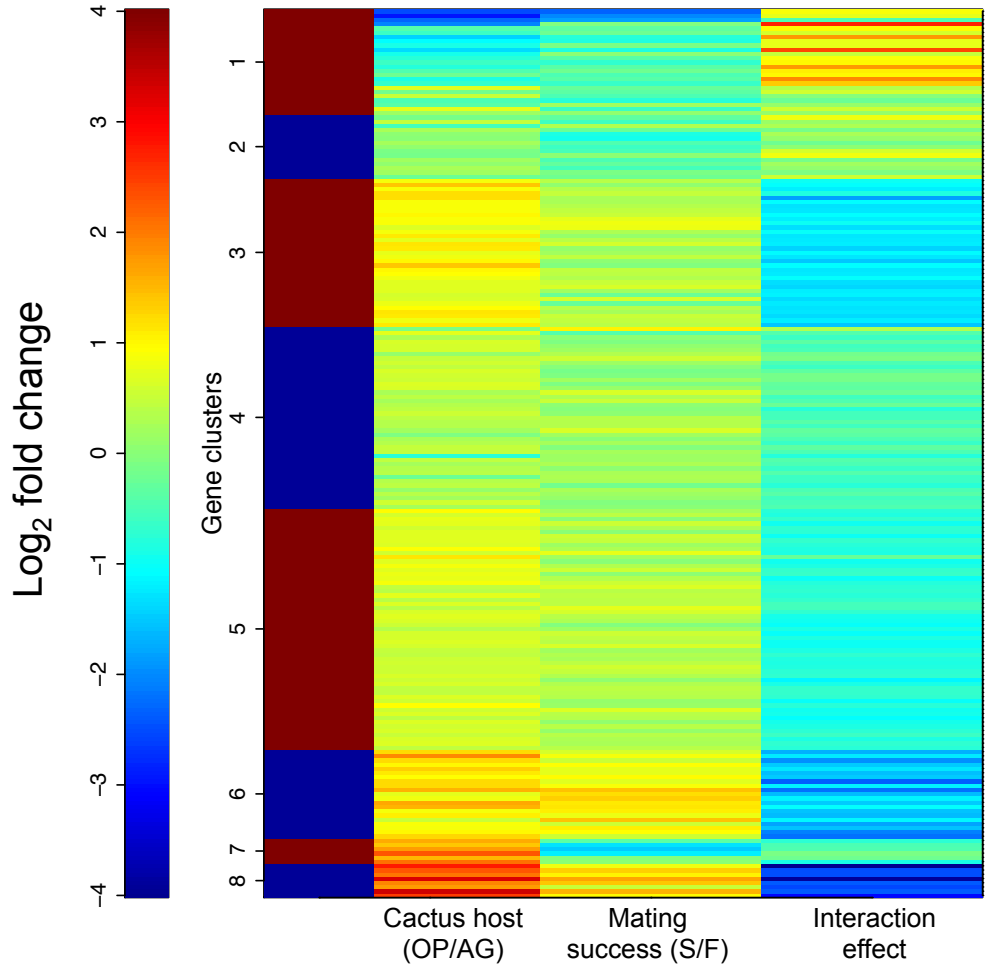


Fig. 4.1. Heat map displaying all differentially expressed genes across the 3 contrasts indicated. The colour key representing the $\log_2$ fold change values for each gene in each contrast is shown to the left, and the grouping of genes into 8 clusters indicated in the main panel. S and F denote successfully and unsuccessfully mated male treatments respectively, OP is the organ pipe cactus host treatment and AG is the agria cactus treatment.

Functional enrichment of DE genes using *D. melanogaster* orthologs produced several significant terms for cactus and the interaction effects, though there was no main effect for mating success. Cactus effects produced a single cluster, containing several different terms involved in immune response (Fig. 4.2). Several terms were individually significant across cacti, though not as a cluster. These included four genes for olfactory behaviour, chemosensory behaviour and cognition (Fig. 4.2; *Obp99A*, *Or83A*, *Gr94A* and *drk*). Two functional clusters were enriched in the interaction between rearing cactus and mating success. The first included terms for ribosome, ribonucleoprotein complex and translation. The second had an enrichment score <1.3, but included significant single annotations for protein targeting to mitochondria.

Several ncRNAs were differentially expressed due to both cactus and interaction effects (Table 4.2). These were small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs). snoRNAs are often located in the introns of genes, particularly those associated with ribosome structure, and are excised from introns by the spliceosome. Alternatively they are transcribed as polycistronic transcripts by RNA poymerase II and processed into multiple RNAs (Terns & Terns 2002). The majority of snoRNAs in Table 4.2 originated from the introns of protein coding genes. Transcription of these genes by RNA polymerase II means that both snRNAs and snoRNAs would have survived size selection from the RNA extraction protocol and poly(A) tail selection, and thus were accurately quantified. These ncRNAs do not have detailed functional annotations and thus were not part of the enrichment analyses. However, of the DE protein-coding genes 14 were identified as known constituents of ribonucleoprotein complexes. Most of these were ribosomal but one, *U2af38*, forms

part of the spliceosome, attaching to the 3' splice site during alternative splicing. Both this and another significantly expressed gene, *LSm-4*, have roles in alternative splicing.

Consistent up-regulation of gene expression in one treatment group over another is often associated with an increased functional importance. The majority of significantly DE genes were up-regulated on organ pipe cactus, the host used in nature in this population, as were genes involved in methylation (Fig. 4.3). The methylation GO term (GO:0032259) involves the attachment of a methyl group to a molecule and is not limited to DNA methylation but includes RNA and protein methylation. Interestingly, the most consistently up-regulated genes in successfully mated males were involved in methylation, the other DE categories surprisingly being down-regulated in successful males. Most glycerol ether metabolism and translation genes were also up-regulated.

Table 4.2 Significantly differentially expressed non-coding RNA products, their orthologs in *D. melanogaster* and their corresponding gene ontology functional annotations.

***D. mojavensis* ID**	***D. melanogaster* ID**	**GO term annotations**
Dmoj\snoRNA:GI25318	Dmel\snoRNA:Psi18S-110	nuclear gene
Dmoj\snoRNA:GI25328	Dmel\snoRNA:Psi28S-2566	nuclear gene; nucleolus
Dmoj\snoRNA:GI25330	Dmel\snoRNA:Psi28S-3327b	nuclear gene
Dmoj\snoRNA:GI25333	Dmel\snoRNA:Me28S-G980	nuclear gene
Dmoj\snoRNA:GI25343	Dmel\snoRNA:Psi18S-841d	nuclear gene
Dmoj\snoRNA:GI25349	Dmel\snoRNA:Psi18S-841a	nuclear gene; nucleolus
Dmoj\snoRNA:GI25354	Dmel\snoRNA:Me28S-C3420a	nuclear gene
Dmoj\snoRNA:GI25358	Dmel\snoRNA:Me28S-G2703a	nuclear gene
Dmoj\snoRNA:GI25368	No orthologous hits	-
Dmoj\snoRNA:GI25378	No orthologous hits	-
Dmoj\snoRNA:GI25382	Dmel\snoRNA:Me18S-C1096	nuclear gene
Dmoj\snoRNA:GI25384	Dmel\snoRNA:Psi28S-2442b	nuclear gene
Dmoj\snoRNA:GI25385	No orthologous hits	-
Dmoj\snoRNA:GI25391	No orthologous hits	-
Dmoj\snoRNA:GI25394	Dmel\snoRNA:Psi18S-1377d	nuclear gene
Dmoj\snoRNA:GI25402	Dmel\snoRNA:Psi28S-3305b	nuclear gene
Dmoj\snoRNA:GI25408	No orthologous hits	-
Dmoj\snoRNA:GI25409	Dmel\snoRNA:Psi28S-1060	nuclear gene
Dmoj\snoRNA:GI25413	Dmel\snoRNA:Psi28S-1135a	nuclear gene
Dmoj\snoRNA:GI25418	Dmel\snoRNA:Psi28S-1135f	nuclear gene
Dmoj\snoRNA:GI25426	Dmel\snoRNA:U14:30Eb	nucleolus; rRNA modification guide activity
Dmoj\snoRNA:GI25427	Dmel\snoRNA:Me18S-A1576	nuclear gene; nucleolus
Dmoj\snoRNA:GI25433	No orthologous hits	-
Dmoj\snoRNA:GI25436	No orthologous hits	-
Dmoj\snRNA:U2:2	Dmel\snRNA:U2:34ABa	U2 snRNP; nuclear mRNA splicing, via spliceosome
Dmoj\snRNA:U4:2	Dmel\snRNA:U4:25F	U4 snRNP; nuclear mRNA splicing, via spliceosome

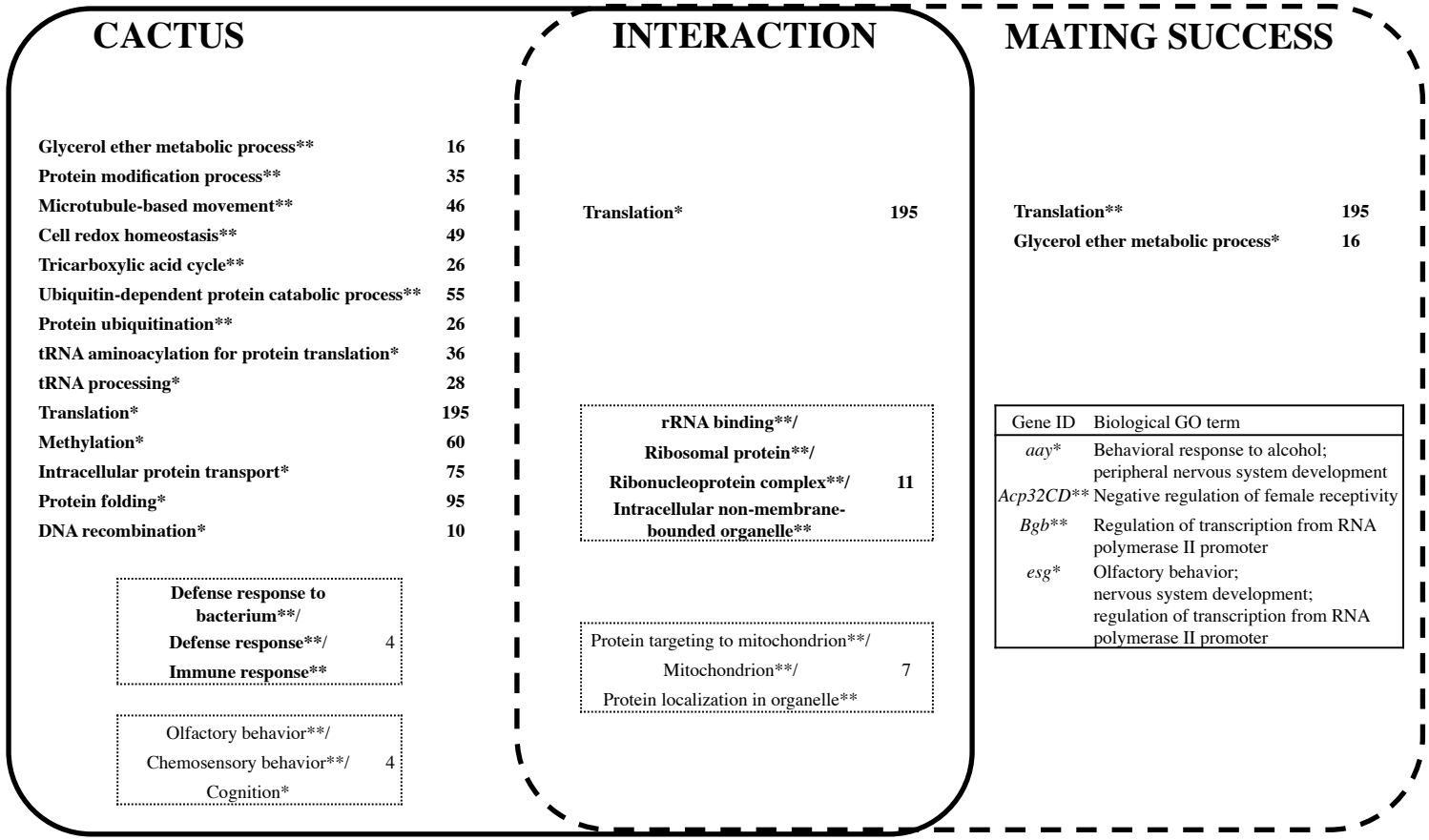


Fig. 4.2 Summary of functional enrichment results for differentially expressed genes. Gene Ontology (GO) terms are grouped by experimental effect; cactus in large solid round box, mating success in large dashed round box, and their interaction. All GO terms are one per line with the number of associated genes indicated. Overlapping and root GO terms were removed for brevity. Rank mean test results are presented as un-boxed, bold GO terms along with significance for each term (**; FDR <0.05 and *; FDR <0.1). DAVID enrichment clusters are shown inside fine dashed boxes; bold and fine dashed boxed terms denoted significant DAVID clusters with an enrichment score of >1.3, fine dashed boxes represent clusters with enrichment score <1.3, yet containing individually significant terms. Significance for each term from DAVID is indicated (**; p<0.05 and *; p<0.1). The solid table presents interesting significant genes (*D. melanogaster* orthologs), and corresponding biological GO terms (**; FDR<0.05 and *; FDR<0.1).

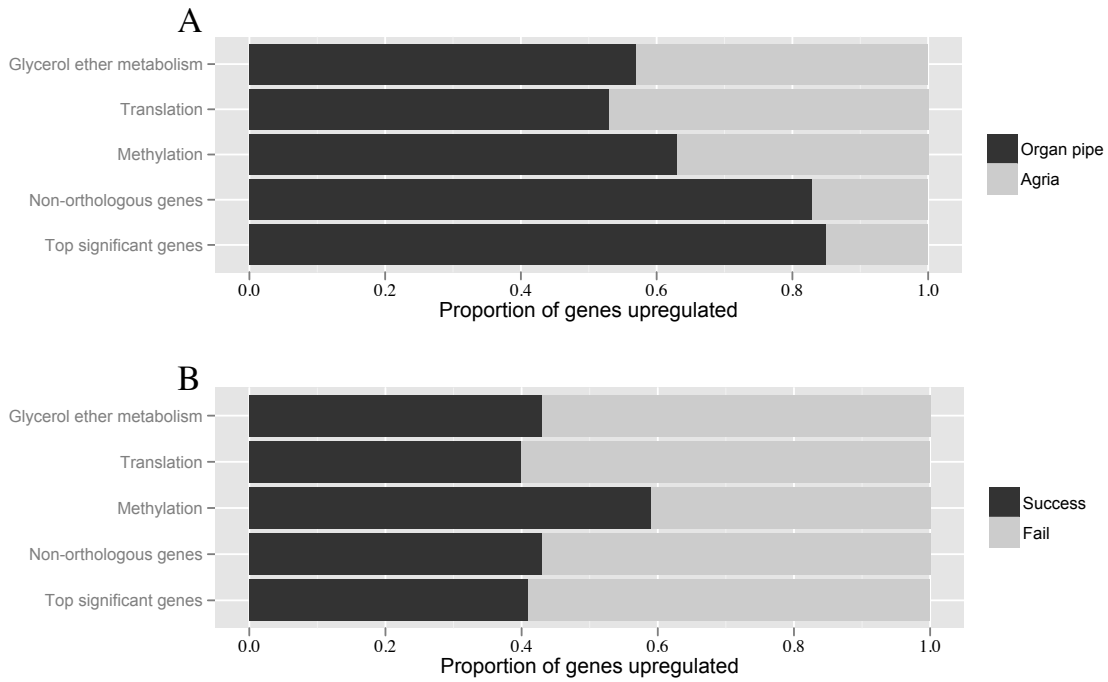


Fig. 4.3 Proportion of genes upregulated on organ pipe cactus (a) and according to mating success (b). Genes are presented by several categories including all significantly differentially expressed genes, non-orthologous genes (i.e. no identified orthologs in *D. melanogaster*), and three different functional categories of genes that were significantly enriched in the rank mean tests.

Alternative expression

Fewer AE genes were detected than DE genes, except in the case of mating success (Table 4.1). Surprisingly, the interaction effect showed only one significant annotation; however, more functions were identified using the most significant *D. melanogaster* orthologs in DAVID.

Rank mean test enrichment suggested functions for cellular signaling and ion transport genes differed across cactus hosts (Fig. 4.4). Annotations included signal transduction and G-protein coupled receptor-signaling pathways that function in the cellular response to extra-cellular signals. The end point of cellular signaling pathways is often the regulation of transcription, and this term was also significantly enriched due to cactus effects. Ion transport and sodium ion transport were seen,

along with calcium ion transport via voltage-gated channels. These terms are often associated with synaptic transmission. Histone deacetylation genes were also AE across cacti, indicating a potential epigenetic response to cactus.

Functional annotations for both transcription, ion transport and chitin metabolic process were seen in the mating success contrasts, along with multicellular organismal development (Fig. 4.4). Lastly, the interaction effect contained only one significant functional annotation, for the regulation of calcium ion transport via voltage-gated channels.

DAVID analysis of AE orthologs revealed a similar set of functional annotations to the rank mean test, with the exception of mating success which showed no significantly enriched GO terms (Fig. 4.4). Cactus-specific genes were enriched for one cluster that included terms for transmission of nerve impulses and synaptic transmission. Other individually significant terms were seen, including alternative splicing. The interaction of cactus and mating success contained two significantly enriched clusters of AE genes. The first included several terms for neuron related development and axonogenesis, consistent with the results of the rank mean test. The second included terms for the regulation of transcription and chromatin modification and regulation. Interestingly, the *slowpoke* (*slo*) gene was AE in both the main cactus and interaction effects. This gene encodes an ion channel protein with biological functions in song production and structure and male courtship behaviour, and is necessary for ethanol tolerance (Cowmeadow *et al*. 2005). It has also been shown to influence male courtship song in *D. melanogaster* (Peixoto & Hall 1998).

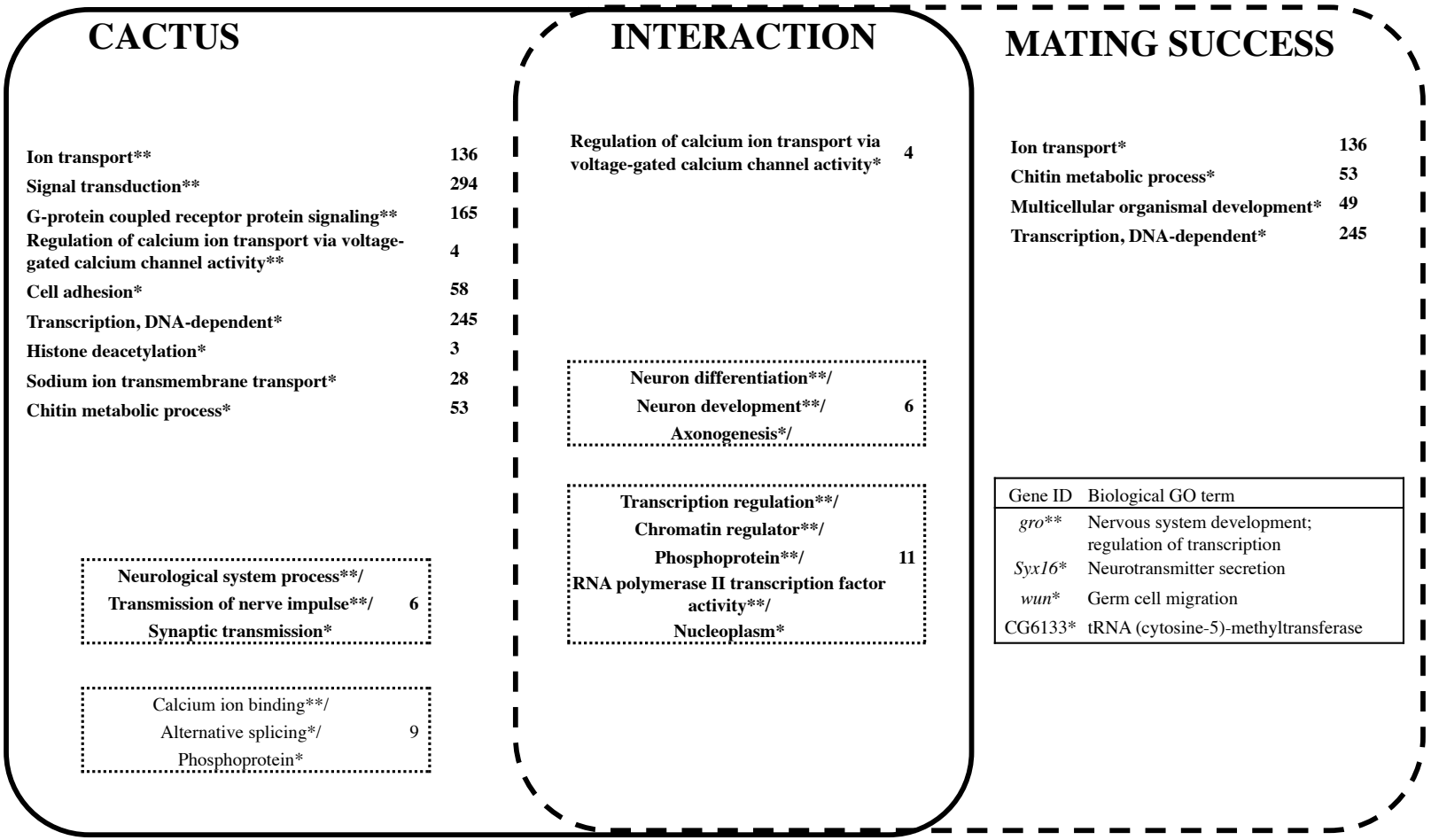


Fig. 4.4 Summary of functional enrichment results for alternatively expressed genes. Figure details are the same as in Fig. 4.2 above.

DISCUSSION

Ecological speciation involves the adaptation of populations to new environments and concurrent evolution of reproductive isolation (Nosil 2012). Functional genetic links between genes involved in ecological adaptation and sexual behaviour have rarely been examined. It is likely that environmental plasticity is common in ecological adaptation and thus traits potentially influencing both adaptation and isolation may often be influenced by coordinated changes in gene expression and plasticity (Thibert-Plante & Hendry, 2011). However, incorporating the analyses of such genes into studies of speciation is in its infancy. Here, using high-throughput sequencing transcriptome sequencing, we distinguished the gene expression changes due to both host plant variation and mating success in cactophilic *D. mojavensis*, and identified the functions of genes involved in cactus-dependent mating success.

The *D. mojavensis* genome is poorly annotated with only 32% of all genes having at least one biological GO annotation, most of which are non-specific (root) terms, and approximately 32% of *D. mojavensis* genes have no known orthologs in *D. melanogaster* (Tweedie *et al*. 2009). In comparison, 72% of *D. melanogaster* genes have been annotated, 67% of which are specific (non-root) annotations (Tweedie *et al*. 2009). Here, approximately two thirds of the significant genes across effects had no functional information (biological process GO terms) and one third of these genes had no discovered orthologs in the *D. melanogaster* genome. Functional enrichment analyses of poorly annotated genomes is thus challenging, especially when the number of significantly differentially expressed genes is small. Consequently a combined approach was taken using two different enrichment methods, to identify functional information. We uncovered clusters of co-regulated gene sets (Fig. 4.1),

and using functional enrichment tests with *D. mojavensis* annotated genes and *D. melanogaster* orthologs produced comparable results on the functions of these genes.

Functional analyses

Cactus-specific differential gene expression showed significant enrichment for GO terms that were also detected in two microarray studies examining cactus-specific gene expression during larval development in *D. mojavensis* (Matzkin *et al*. 2006; Matzkin 2012). These included functions in immune response, metabolism, signal transduction and the nervous system. Previous work has also shown that the transcriptomic response to cacti in larvae, and to desiccation, involve key metabolic pathways, including the TCA cycle (Matzkin & Markow 2009). Here, DE genes were seen that function in chemical metabolism, including glycerol ether metabolism, which aids in the assimilation of volatile alcohols. *D. mojavensis* is able to metabolise ethanol vapour, with consequent effects on life history traits such as longevity, life time fecundity and metabolic rates (Starmer *et al*. 1977; Etges & Klassen 1989). Cactus-specific effects also included genes involved in the TCA cycle and cell redox homeostasis. Cellular oxidative stress is known to accelerate cellular damage and shorten lifespan in *Drosophila* (Ruan *et al*. 2002), and is often linked to changes in metabolism. We studied sexually mature flies and found that the preadult rearing environment therefore has a carry-over effect onto adult gene expression functioning in chemical metabolism. The chemical environment of columnar cacti is well documented. Alkaloids, medium chain fatty acids, sterol diols and triterpene glycosides have all been shown to be largely species-specific for Sonoran Desert cacti, and are causal factors in explaining patterns of host plant use in cactophilic *Drosophila* (Fogleman & Danielson 2001). *D. mojavensis* is oligophagous due to its

ability to metabolise medium chain fatty acids, sterol diols and high levels of triterpene glycosides found in organ pipe and agria cacti (Fogleman & Danielson 2001). Organ pipe and agria differ in their triterpene glycoside content with agria containing higher levels than organ pipe cactus, however it is not known if these compounds are differentially metabolised in Baja California vs mainland populations of *D. mojavensis*.

The fermenting cactus environment also caused differential expression of genes that function in protein modification, specifically protein ubiquitination in protein catabolism. Ubiquitination of proteins marks them for degradation by proteasomes and regulates protein levels for a host of critical cellular functions, including gene expression regulation (Pickart 2001; Shilatifard 2006). Ubiquitination also plays a role in stress and immune system responses and the latter term was significantly enriched under the rank mean test (Fig. 4.2).

Ecological links to speciation would predict a connection between adaptation to cactus and mating success, but relatively few enriched functional groups were found to differ between males who were successful and unsuccessful in mating, either directly or in interaction with cactus, regardless of the number of DE genes. Only one broad term, translation, was strongly enriched for the interaction effect using the rank mean test. Functional clustering of these genes showing interaction effects produced terms for ribosome function, ribonucleoprotein complex and translation. These genes also included several ncRNAs and splicing factors. Such ncRNAs are involved in the production of mature messenger RNA, RNA modifications and translation. They also included snRNAs, which are the backbone of the spliceosome, and snoRNAs, which modify snRNAs, rRNA and mRNA (Kiss 2002). Further, it has been shown that snRNAs and snoRNAs can function in pre-mRNA processing through involvement in

splice site selection (Kishore & Stamm 2006; Matera *et al*. 2007; Khanna & Stamm 2010). Two splicing factors were significantly DE, orthologous to *U2af38* and *LSm-4* in *D. melanogaster*, which regulate alternative splicing (Park *et al*. 2004; Tritschler *et al*. 2007). *U2af38* is a core splicing component that forms part of the spliceosome, attaching to the 3' splice site during alternative splicing and Sm-like proteins such as *LSm-4* associate with small RNA components of the spliceosome, influencing the alternative expression of genes (Will & Lührmann 2011).

Alternatively expressed genes had roles in cellular signaling, neurological development, gene expression regulation and organismal development. Cactus-specific functional enrichment implicated a role for intracellular signaling in response to extracellular cues. G-protein coupled receptor signaling pathways are a large family of cell surface molecules that act as receptors for a range of stimuli, including neurotransmitters, hormones, growth factors, odorant molecules and light (Marinissen & Gutkind 2001). Other enriched functions among AE cactus related genes included terms for the transmission of nerve impulses. These included calcium and sodium ion transport through voltage-gated channels, and synaptic transmission itself. Thus, rearing flies on differing hosts until eclosion had lasting effects on the expression of genes involved in extracellular signaling and transmission of nerve impulses and possibly adult behaviour.

Links to behavioural phenotypes

A potential link between larval cactus effects and adult behavioural phenotypes was discovered through the DE and AE of behaviour related genes. Several chemosensory behaviour genes were differentially expressed across host plants, most likely relating to chemical differences between cacti. Such behaviour includes sensing of volatile

chemicals for host plant detection (Fogleman & Danielson 2001), and evidence suggests that gustatory receptor genes, such as *Gr94A*, have neuronal links to the reproductive organs in *Drosophila* (Park & Kwon 2011). The AE of nervous system related genes suggests that alternative exon use might be particularly important for behavioural plasticity. However, 10 significantly DE genes in the interaction effect were also annotated with neurogenesis/nervous system development functions. The expression of chemosensory and impulse transmission genes indicates cactus-specific differences in sensing and parsing of cues from the external chemical environment. This, in turn, has a potential influence on adult male mating success, through cactus-specific expression of nervous system development genes. The peripheral and central nervous systems both play a major role in *Drosophila* mating behaviour (Villella & Hall 2008). Mutations in *Drosophila* ion channel genes are known to influence behaviour, such as learning and olfaction, as well as courtship song production (Peixoto & Hall 1998; Gleason 2005). Ion channel genes are therefore good candidates for controlling mating behavior (Kyriacou 2002) and demonstrate complex expressional regulation through alternative splicing (Smith *et al*. 1996). Expression variation in ion channel and nervous system development genes such as *slo*, which we found to be significantly AE, may have an influence on courtship behaviour in flies. Further evidence of a role for *slowpoke* in *D. mojavensis* evolution comes from a QTL study in which *slo* was identified as a potential candidate gene underlying courtship song production (Etges *et al*. 2007). Thus the expression of chemosensory and nervous system related genes across cacti and mating success treatments provide a potential link between cactus hosts and adult courtship behaviour, through nervous system development.

QTL studies have shown strong GxE effects between host cactus and traits involved in mating behavior in *D. mojavensis*, through differences in epicuticular hydrocarbons (used as contact pheromones) and courtship song production (Etges *et al*. 2007; Etges *et al*. 2009; Etges *et al*. 2010). Surprisingly few genes directly related to CHC production were seen here, although genes involved in metabolism, seen in both cactus and mating effects, might also play a role. Genes such as the Δ9 desaturases are thought to be important in pheromone production (Keays *et al*. 2011), were implicated in *D. mojavensis* QTL studies and may influence reproductive isolation between *D. melanogaster* populations (Dallerac *et al*. 2000; Takahashi *et al*. 2001). However, no desaturase genes were significantly differentially or alternatively expressed in these analyses perhaps suggesting that any GxE effects of these genes do not involve an expression response. This suggests there may be a greater role for plasticity of neurological function, potentially involving courtship song, and chemosensory behaviours in *D. mojavensis*.

Gene expression changes over the life cycle

We examined the effect of larval environmental manipulation on adult gene expression (rather than cross-generational effects) and found significant functional enrichment of several different types of epigenetic modifications. Because flies were raised on cactus hosts only during egg to eclosion, cactus specific gene expression patterns may have been laid down during this period, and propagated through to adulthood. Several processes that can play such a role were identified. Cactus-specific expression included genes functioning in methylation (e.g. ortholog of *pr-set7 D. melanogaster* gene), protein ubiquitination and histone modification (e.g. orthologs of *Snp* and CG31703). The significant methylation GO term broadly includes any

attachment of a methyl group to protein, DNA or RNA. Interestingly a higher proportion of methylation related genes were upregulated in successfully mated males and most other DE genes were downregulated (Fig. 4.3). Increased methylation is thought to repress gene expression (Wolffe & Matzke 1999), meaning that mating success might be particularly influenced by methylation-based control of gene regulation.

Epigenetic modifications of RNA and proteins rather than DNA might be particularly important in *Drosophila* as this group does not have the full complement of DNA methyltransferases commonly found in other organisms, having only retained one methytransferase, orthologous to the human *Dnmt2* gene (Lyko & Maleszka 2011). *Dnmt2* is only thought to function in the methylation of tRNA (although this has been disputed; Goll *et al*. 2006; Krauss & Reuter 2011), meaning that the methylation GO term seen in this study will mainly involve RNA and proteins. tRNA processing was a significantly enriched term in the rank mean test of DE genes, and a tRNA methyltransferase, *Nsun2*, was significantly AE. *Nsun2* functions in spermatogenesis in *Drosophila* (Gerbasi *et al*. 2011) and splicing mutations within it can cause short-term memory loss, demonstrating the importance of correct splicing for function of this gene (Abbasi-Moheb *et al*. 2012). Alternative splicing has also been suggested as an important mechanism underlying phenotypic plasticity (Marden 2008) and histone modifications and chromatin remodeling are known to play a role in alternative splicing (Luco *et al*. 2010; Luco *et al*. 2011). Chromatin remodeling genes are important for temperature related plasticity in *D. melanogaster* (Levine *et al*. 2011) and chromatin assembly genes were differentially expressed in cactus-specific larval plasticity in *D. mojavensis* (Matzkin *et al*, 2006).

Evidence for a suite of epigenetic processes associated with host plant plasticity suggest that *Drosophila* employ gene regulatory mechanisms other than DNA methylation, and that these mechanisms might play a role in ecological adaptation. There has been some controversy surrounding the role of epigenetic mechanisms in *Drosophila* species, specifically whether DNA methylation routinely occurs in the *Drosophila* genome. Recent studies suggest that methylation does occur (Krauss & Reuter 2011), yet experimental evidence indicates this is at low levels genome-wide, and the functional significance remains unclear (Lyko *et al*. 2000). Our results suggest a potential role for RNA and protein methylation that links larval cactus plasticity with adult phenotypes. Recent evidence suggests an important role of chromatin modification in gene expression plasticity in *Drosophila* (Levine *et al*. 2011). Therefore, species lacking the core *Dnmt* genes, the '*Dnmt2* only' species (Krauss & Reuter 2011), might regulate their genome through mechanisms other than, or in addition to, DNA methylation. This might include RNA and histone protein modifications, and involve snRNAs and snoRNAs.

Potential links between ecological adaptation and reproductive success have rarely been addressed at a transcriptomic level. Agria cactus causes decreased mate discrimination and higher mating success, particularly for Baja California males, in multiple choice studies (Etges 1992). QTLs for mating success and the phenotypes involved often show genotype-by-environment interactions (Etges *et al*. 2007; Etges *et al*. 2009). Few DE or AE genes were seen to be present in both the cactus and mating success main effects, indicating little evidence for a shared genetic basis, or pleiotropy. However, models of ecological speciation include other modes of linking ecological adaptation and reproductive isolation, such as physical linkage (Rundle & Schluter 2004). Here we found that larval host cactus influenced adult male mating

success by modulating the expression of genes involved in translation, transcription and nervous system development. Gustatory receptor genes such as *Gr94A* and nervous system genes such as *slo* have been linked to reproduction and courtship behaviour in *Drosophila* (Peixoto & Hall 1998; Park & Kwon 2011). The expression of such genes here suggests that the genetic basis of mating success in *D. mojavensis* is likely to involve nervous system development genes that link the cactus environment to reproductive behaviour. Mainland populations diverged from an ancestral Baja California population around 230-270,000 years ago (Smith *et al*. 2012). This suggests that the adaptation of *D. mojavensis* to organ pipe cactus and the concurrent evolution of reproductive isolation has been fairly rapid, and that plasticity in gene expression may have played an important role in this. Examining the molecular architecture that underlies plasticity of gene expression is therefore an important step towards understanding the role of gene expression in ecological speciation (Pavey *et al*. 2010).

REFERENCES

Abbasi-Moheb L, Mertel S, Gonsior M, *et al*. (2012) Mutations in NSUN2 cause autosomal-recessive intellectual disability. *American Journal of Human Genetics*, **90**, 847-855.

Alcorn S, Orum T, Steigerwalt AG, *et al*. (1991) Taxonomy and pathogenicity of *Erwinia cacticida* sp. nov. *International Journal of Systematic Bacteriology*, **41**, 197-212.

Altschul SF, Madden TL, Schaffer AA, *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

Anders (2010) Htseq: Analysing high-throughput sequencing data with python. [http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html].

Binns D, Dimmer E, Huntley R, *et al*. (2009) QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics*, **25**, 3045-3046.

Butlin R, Debelle A, Kerth C, *et al*. (2012) What do we need to know about speciation? *Trends in Ecology & Evolution*, **27**, 27-39.

Clark AG, Eisen MB, Smith DR, *et al*. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**, 203-218.

Cowmeadow RB, Krishnan HR, Atkinson NS (2005) The slowpoke gene is necessary for rapid ethanol tolerance in *Drosophila*. *Alcoholism: Clinical and Experimental Research*, **29**, 1777-1786.

Crispo E (2008) Modifying effects of phenotypic plasticity on interactions among natural selection, adaptation and gene flow. *Journal of Evolutionary Biology*, **21**, 1460-1469.

Dallerac R, Labeur C, Jallon JM, *et al*. (2000) A Delta-9 desaturase gene with a different substrate specificity is responsible for the cuticular diene hydrocarbon polymorphism in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences, USA*, **97**, 9449-9454.

Dennis JG, Sherman BT, Hosack DA, *et al*. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology*, **4**, R60-R60.11.

Etges WJ (1992) Premating isolation is determined by larval substrates in cactophilic *Drosophila mojavensis*. *Evolution*, **46**, 1945-1950.

Etges WJ (1998) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. IV. Correlated responses in behavioral isolation to artificial selection on a life-history trait. *American Naturalist*, **152**, 129-144.

Etges WJ, Ahrens MA (2001) Premating isolation is determined by larval-rearing substrates in cactophilic *Drosophila mojavensis*. V. Deep geographic variation in epicuticular hydrocarbons among isolated populations. *American Naturalist*, **158**, 585-598.

Etges WJ, de Oliveira CC, Gragg E, *et al*. (2007) Genetics of incipient speciation in *Drosophila mojavensis*. I. Male courtship song, mating success, and genotype x environment interactions. *Evolution*, **61**, 1106-1119.

Etges WJ, De Oliveira CC, Noor MAF, Ritchie MG (2010) Genetics of incipient speciation in *Drosophila mojavensis*. III. Life history divergence in allopatry and reproductive isolation. *Evolution*, **64**, 3549-3569.

Etges WJ, de Oliveira CC, Ritchie MG, Noor MAF (2009) Genetics of incipient speciation in *Drosophila mojavensis*. II. Host plants and mating status

influence cuticular hydrocarbon QTL expression and G× E interactions. *Evolution*, **63**, 1712-1730.

Etges WJ, Johnson WR, Duncan GA, Huckins G, Heed WB (1999) Ecological genetics of cactophilic *Drosophila*. In: *Ecology of Sonoran Desert plants and plant communities*. (ed. Robichaux, R), pp. 164-214. University of Arizona Press, Tucson.

Etges WJ, Klassen CS (1989) Influences of atmospheric ethanol on adult *Drosophila mojavensis*: Altered metabolic rates and increases in fitness among populations. *Physiological Zoology*, **62**, 170-193.

Fogleman JC, Armstrong L (1989) Ecological aspects of cactus triterpene glycosides .1. Their effect on fitness components of *Drosophila mojavensis*. *Journal of Chemical Ecology*, **15**, 663-676.

Fogleman JC, Danielson PB (2001) Chemical interactions in the cactus-microorganism-*Drosophila* model system of the Sonoran Desert. *American Zoologist*, **41**, 877-889.

Fogleman JC, Starmer WT, Heed WB (1981) Larval selectivity for yeast species by *Drosophila mojavensis* in natural substrates. *Proceedings of the National Academy of Sciences, USA*, **78**, 4435-4439.

Funk DJ, Nosil P, Etges WJ (2006) Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. *Proceedings of the National Academy of Sciences, USA*, **103**, 3209-3213.

Gerbasi VR, Preall JB, Golden DE, *et al*. (2011) Blanks, a nuclear siRNA/dsRNA-binding complex component, is required for *Drosophila* spermiogenesis. *Proceedings of the National Academy of Sciences, USA*, **108**, 3204-3209.

Ghalambor C, McKay J, Carroll S, Reznick D (2007) Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Functional Ecology*, **21**, 394-407.

Gleason JM (2005) Mutations and natural genetic variation in the courtship song of *Drosophila*. *Behavior Genetics*, **35**, 265-277.

Goll MG, Kirpekar F, Maggert KA, *et al*. (2006) Methylation of tRNAAsp by the DNA methyltransferase homolog *Dnmt2*. *Science*, **311**, 395-398.

Griffith M, Griffith OL, Mwenifumbo J, *et al*. (2010) Alternative expression analysis by RNA sequencing. *Nature Methods*, **7**, 843-U108.

Hartigan JA, Wong MA (1979) Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **28**, 100-108.

Havens JA, Orzack SH, Etges WJ (2011) Mate choice opportunity leads to shorter offspring development time in a desert insect. *Journal of Evolutionary Biology*, **24**, 1317-1324.

Heed WB (1982) The origin of *Drosophila* in the Sonoran Desert. In: *Ecological genetics and evolution: The cactus-yeast-Drosophila model system*. (Barker, JSF and Starmer, WT eds.), Pp. 65-80. Academic Press, Sydney.

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**, 44-57.

Keays MC, Barker D, Wicker-Thomas C, Ritchie MG (2011) Signatures of selection and sex-specific expression variation of a novel duplicate during the evolution of the *Drosophila* desaturase gene family. *Molecular Ecology*, **20**, 3617-3630.

Khanna A, Stamm S (2010) Regulation of alternative splicing by short non-coding nuclear RNAs. *RNA Biology*, **7**, 480-485.

Kim S, Kim H, Fong N, Erickson B, Bentley DL (2011) Pre-mRNA splicing is a determinant of histone H3K36 methylation. *Proceedings of the National Academy of Sciences, USA*, **108**, 13564-13569.

Kishore S, Stamm S (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, **311**, 230-232 .

Kiss T (2002) Small nucleolar RNAs: An abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145-148.

Krauss V, Reuter G (2011) DNA methylation in *Drosophila*: A critical evaluation. *Progress in Molecular Biology and Translational Science*, **101**, 177-191.

Levine MT, Eckert ML, Begun DJ (2011) Whole-genome expression plasticity across tropical and temperate *Drosophila melanogaster* populations from Eastern Australia. *Molecular Biology and Evolution*, **28**, 249-256.

Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T (2011) Epigenetics in alternative pre-mRNA splicing. *Cell*, **144**, 16-26.

Luco RF, Pan Q, Tominaga K, *et al*. (2010) Regulation of alternative splicing by histone modifications. *Science*, **327**, 996-1000.

Lyko F, Maleszka R (2011) Insects as innovative models for functional studies of DNA methylation. *Trends in Genetics*, **27**, 127-131.

Lyko F, Ramsahoye BH, Jaenisch R (2000) DNA methylation in *Drosophila melanogaster*. *Nature*, **408**, 538-540.

Maan ME, Seehausen O (2011) Ecology, sexual selection and speciation. *Ecology Letters*, **14**, 591-602.

Machado CA, Matzkin LM, Reed LK, Markow TA (2007) Multilocus nuclear sequences reveal intra and interspecific relationships among chromosomally polymorphic species of cactophilic *Drosophila*. *Molecular Ecology*, **16**, 3009-3024.

Marden JH (2008) Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. *Heredity*, **100**, 111-120.

Marinissen MJ, Gutkind JS (2001) G-protein-coupled receptors and signaling networks: emerging paradigms. *Trends in Pharmacological Sciences*, **22**, 368-376.

Markow TA, Fogleman JC, Heed WB (1983) Reproductive Isolation in Sonoran Desert *Drosophila*. *Evolution*, **37**, 649-652.

Markow TA (1991) Sexual isolation among populations of *Drosophila mojavensis*. *Evolution*, **45**, 1525-1529.

Matera AG, Terns RM, Terns MP (2007) Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature Reviews: Molecular Cell Biology*, **8**, 209-220.

Matzkin LM (2004) Population genetics and geographic variation of alcohol dehydrogenase (*Adh*) paralogs and glucose-6-phosphate dehydrogenase (*G6pd*) in *Drosophila mojavensis*. *Molecular Biology and Evolution*, **21**, 276-285.

Matzkin LM (2012) Population transcriptomics of cactus host shifts in *Drosophila mojavensis*. *Molecular Ecology*, **21**, 2428-2439.

Matzkin LM, Eanes WF (2003) Sequence variation of alcohol dehydrogenase (*Adh*) paralogs in cactophilic *Drosophila*. *Genetics*, **163**, 181-194.

Matzkin, LM, Markow TA (2009) Transcriptional regulation of metabolism associated with the increased desiccation resistance of the cactophilic *Drosophila mojavensis*. *Genetics*, **182**,1279-1288.

Matzkin LM, Watts TD, Bitler BG, Machado CA, Markow TA (2006) Functional genomics of cactus host shifts in *Drosophila mojavensis*. *Molecular Ecology*, **15**, 4635-4643.

McQuilton P, St Pierre SE, Thurmond J, Consortium F (2012) FlyBase 101- the basics of navigating FlyBase. *Nucleic Acids Research*, **40**, D706-D714.

Nosil P (2012) *Ecological Speciation*. Oxford University Press, Oxford, UK.

Park JW, Parisky K, Celotto AM, Reenan RA, Graveley BR (2004) Identification of alternative splicing regulators by RNA interference in *Drosophila*. *Proceedings of the National Academy of Sciences, USA*, **101**, 15974-15979.

Park JH, Kwon JY (2011) A systematic analysis of *Drosophila* gustatory receptor gene expression in abdominal neurons which project to the central nervous system. *Molecules and Cells*, **32**, 375-381.

Pavey SA, Collin H, Nosil P, Rogers SM (2010) The role of gene expression in ecological speciation. *Annals of the New York Academy of Sciences*, **1206**, 110-129.

Peixoto AA, Hall JC (1998) Analysis of temperature-sensitive mutants reveals new genes involved in the courtship song of *Drosophila*. *Genetics*, **148**, 827-838.

Pfennig DW, Wund MA, Snell-Rood EC, *et al*. (2010) Phenotypic plasticity's impacts on diversification and speciation. *Trends in Ecology & Evolution*, **25**, 459-467.

Pickart CM (2001) Mechanisms underlying ubiquitination. *Annual Review of Biochemistry*, **70**, 503-533.

Rajpurohit S, de Oliveira CC, Etges WJ, Gibbs AG (2013) Functional genomic and phenotypic responses to desiccation in natural populations of a desert drosophilid. *Molecular Ecology: In Press*.

Reed LK, Nyboer M, Markow TA (2007) Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. *Molecular Ecology*, **16**, 1007-1022.

Ritchie MG (2007) Sexual selection and speciation. *Annual Review of Ecology, Evolution and Systematics*, **38**, 79-102.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.

Ross C, Markow T (2006) Microsatellite variation among diverging populations of *Drosophila mojavensis*. *Journal of Evolutionary Biology*, **19**, 1691-1700.

Ruan H, Tang XD, Chen ML, *et al*. (2002) High-quality life extension by the enzyme peptide methionine sulfoxide reductase. *Proceedings of the National Academy of Sciences, USA*, **99**, 2748-2753.

Ruiz A, Heed W, Wasserman M (1990) Evolution of the mojavensis cluster of cactophilic *Drosophila* with descriptions of two new species. *Journal of Heredity*, **81**, 30-42.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336-352.

Rundle HD, Schluter D (2004) Natural selection and ecological speciation in sticklebacks. In: *Adaptive Speciation*. (UD, MD, JM, DT, eds.), Pp. 192-209. International Institute for Applied Systems Analysis. Cambridge University Press, UK.

Schluter D (2001) Ecology and the origin of species. *Trends in Ecology & Evolution*, **16**, 372-380.

Scoville AG, Pfrender ME (2010) Phenotypic plasticity facilitates recurrent rapid adaptation to introduced predators. *Proceedings of the National Academy of Sciences, USA*, **107**, 4260-4263.

Servedio MR, Van Doorn GS, Kopp M, Frame AM, Nosil P (2011) Magic traits in speciation: 'magic' but not rare? *Trends in Ecology & Evolution*, **26**, 389-397.

Shilatifard A (2006) Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. *Annual Review of Biochemistry*, **75**, 243-269.

Smadja CM, Butlin RK (2011) A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, **20**, 5123-5140.

Smith G, Lohse K, Etges WJ, Ritchie MG (2012) Model-based comparisons of phylogeographic scenarios resolve the intraspecific divergence of cactophilic *Drosophila mojavensis*. *Molecular Ecology*, **21**, 3293–3307.

Smith LA, Wang X, Peixoto AA, Neumann EK, Hall LM, Hall JC (1996) A *Drosophila* calcium channel alpha1 subunit gene maps to a genetic locus associated with behavioral and visual defects. *Journal of Neuroscience*, **16**, 7868-7879.

Sobel JM, Chen GF, Watt LR, Schemske DW (2010) The biology of speciation. *Evolution*, **64**, 295-315.

Starmer WT (1982) Analysis of the community structure of yeasts associated with the decaying stems of cactus. I. *Stenocereus gummosus*. *Microbial Ecology*, **8**, 71-81.

Starmer WT, Heed WB, Rockwood-Sluss E (1977) Extension of longevity in *Drosophila mojavensis* by environmental ethanol: differences between subraces. *Proceedings of the National Academy of Sciences, USA*, **74**, 387-391.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440-9445.

Takahashi A, Tsaur SC, Coyne JA, Wu C-I (2001) The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences, USA*, **98**, 3920-3925.

Terns MP, Terns RM (2002) Small nucleolar RNAs: Versatile trans-acting molecules of ancient evolutionary origin. *Gene Expression*, **10**, 17-39.

Thibert-Plante X, Hendry A (2010) The consequences of phenotypic plasticity for ecological speciation. *Journal of Evolutionary Biology*, **24**, 326-342.

Tritschler F, Eulalio A, Truffault V, *et al*. (2007) A divergent Sm fold in EDC3 proteins mediates DCP1 binding and P-body targeting. *Molecular and Cellular Biology*, **27**, 8600-8611.

Tweedie S, Ashburner M, Falls K, *et al*. (2009) FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Research*, **37**, D555-D559.

Van Doorn GS, Edelaar P, Weissing FJ (2009) On the origin of species by natural and sexual selection. *Science*, **326**, 1704-1707.

Villella A, Hall JC (2008) Neurogenetics of courtship and mating in *Drosophila*. *Advances in Genetics*, **62**, 67-184.

Wasserman M (1992) Cytological evolution of the *Drosophila repleta* species group. In: *Inversion polymorphism in Drosophila*. (eds. PJR, KCB), pp. 455-541. CRC Press Inc., Bota Racon, Florida.

West-Eberhard MJ (2005a) Developmental plasticity and the origin of species differences. *Proceedings of the National Academy of Sciences, USA*, **102**, 6543-6549.

West-Eberhard MJ (2005b) Phenotypic accommodation: adaptive innovation due to developmental plasticity. *Journal of Experimental Zoology, Part B: Molecular and Developmental Evolution*, **304**, 610-618.

Wolffe AP, Matzke MA (1999) Epigenetics: Regulation through repression. *Science*, **286**, 481-486.

CHAPTER 5:

EVIDENCE FOR EPIGENETIC IMPRINTING IN THE *INSULIN-LIKE GROWTH FACTOR-2* (*IGF2*) GENE OF *GIRARDINICHTHYS MULTIRADIATUS*

ABSTRACT

The relative contribution of natural and sexual selection to speciation is still under debate. Sexual conflict might be play a role in divergence as it can drive the antagonistic coevolution of traits between the sexes. One outcome of sexual conflict is parent-of-origin effects that can occur through gene expression regulation, controlled through genomic imprinting by DNA methylation. Genomic imprinting in the *Insulin-like Growth Factor II* (*Igf2*) gene has been linked to sexual conflict in mammals, where allele-specific imprints regulate gene expression leading to changes in developmental growth. Recent studies in a freshwater Goodeid fish, *Girardinichthys multiradiatus*, show that offspring body size is dependent on the population-of-origin of each parent. Using sequence information from a recent study of *Igf2*, this chapter looked for evidence of genomic imprinting between populations of *G. multiradiatus*. The majority of the *Igf2* sequence was obtained in *G. multiradiatus*, in order to identify potential regulatory features and design primers to examine DNA methylation patterns across the gene. Sequence information from intron 1 and intron 3 was obtained, along with some exonic sequence. Intron 2 did not amplify. Highly conserved 200bp regions were discovered in both sequenced introns, associated with a methylated CpG island in intron 3. This potential regulatory feature might represent a mechanism of genomic imprinting common to all fish. Little evidence for allele-specific methylation was seen, except for one CpG position located immediately upstream from the putative regulatory region. A protocol for direct sequencing of bisulfite converted PCR products was also developed, that allows for the sequencing of multiple samples in order to examine DNA methylation patterns on a population scale and statistically examine methylation variation between treatment groups.

INTRODUCTION

Speciation proceeds with the divergence of populations and the evolution of reproductive isolation (Coyne & Orr 2004). Both natural and sexual selection are thought to play a role in speciation, yet the relative contribution of each is still unclear. Although conceptually it is easy to see how selection on reproductive traits might cause reproductive isolation, studies suggest that sexual selection alone may not be enough to produce a stable divergence of populations (Maan & Seehausen 2011). However, recent theoretical work does suggest that even a very weak source of divergent or disruptive selection, due to environmental heterogeneity, might assist sexual selection in driving divergence (M'Gonigle et al, 2012). Further, natural and sexual selection acting in concert are likely to be a strong driver of speciation, even in the presence of extensive gene flow (Van Doorn *et al*. 2009). The potential link between ecology and reproductive isolation in driving speciation has been supported empirically (Funk *et al*. 2006), and thus sexual selection is likely to play an integral role in the creation of biodiversity.

Some evidence for a role of sexual conflict in assortative mating has been seen, in the absence of environmental selection (Martin & Hosken 2003). Sexual conflict involves the sexually antagonistic coevolution of traits between sexes. This differs from traditional sexual selection, which involves convergent interests or the exploitation of one sex, with sexual conflict being driven by divergent evolutionary interests and the avoidance of costs imposed by the opposite sex (Chapman *et al*. 2003). Conflicts can arise between males and females when each sex attempts to maximize their fitness in incompatible ways, often leading to interlocus and/or intralocus conflict. Interlocus conflict occurs when a trait evolves to enhance the fitness of one sex, with a cost to the other. This may drive evolution in the opposite

sex of 'defensive' traits, creating an arms race of coevolving traits and genetic loci (Chapman *et al*. 2003). Intralocus conflicts arise from divergent selection pressures between sexes on a shared trait controlled by one locus (Bonduriansky & Chenoweth 2009). Conflicts between males and females can occur for many different types of trait and the possible outcomes of sexual conflicts can be diverse (see Gavrilets & Hayashi 2005). These outcomes can include sexual dimorphism or exaggerated reproductive traits, however antagonistic conflicts may have an upper threshold beyond which there is evolution towards resolution, or a mitigation of antagonistic effects (Chapman 2006). Theoretical models examining the role of sexual conflict in speciation demonstrate the potential for the rapidly coevolving male and female traits to lead to the evolution of reproductive isolation, and thus speciation, in both allopatry and sympatry (Gavrilets & Hayashi 2005; Gavrilets & Waxman 2002; Parker & Partridge 1998). However, although sexual conflict may result in genetic diversification it may not necessarily lead to reproductive isolation, and conflict may actually inhibit the process of speciation (Kirkpatrick & Nuismer 2004; Ritchie 2007). One interesting example of the potential results of sexual conflict comes from the Goodinae group of viviparous fish where males exploit females for mating using bright yellow tail markings, which female fish mistake for prey (Macías Garcia & Ramirez 2005). Females pay a cost in foraging as well as in suboptimal mating by reacting to this cue. However, in some species this conflict has lead to females evolving separate sexual and feeding responses, with the male cue evolving into an honest signal.

One potential outcome of sexual conflict is the evolution of parent-of-origin effects, which occur through imprinting and for which there are well characterised mechanisms in mammals (Murphy & Jirtle 2003). Conflicts can arise between males

and females over the level of parental investment (Chapman *et al*. 2003). For example, females might provision their offspring in a manner that maximizes their own fitness. Males might then respond to such maternal effects through imprinting of alleles at particular loci that manipulate offspring fitness towards their own optimum. Imprinting is the differential expression of alleles depending on the parent from which it originated (Brandvain *et al*. 2011; Haig 2000; Haig & Westoby 1989). Evidence suggests that a large number of loci are differentially expressed due to imprinting (Brandvain *et al*. 2011), with approximately 800 genes in mice demonstrating parent-of-origin expression differences in the brain alone (Gregg *et al*. 2010a; Gregg *et al*. 2010b). Imprinting is phylogenetically wide spread, occurring in insects, mammals, fish and plants (Bartolomei & Tilghman 1997; Garnier *et al*. 2008; Lloyd *et al*. 1999; Martin & Mcgowan 1995). The molecular control of genomic imprinting must necessarily be dynamic because imprinting of genes can be erased and reestablished in the gametes each generation (Bartolomei & Tilghman 1997). Epigenetic marks provide just such a dynamic control of gene expression, encompassing a range of protein, DNA and RNA modifications (Berger *et al*. 2009; Bird 2007; Johannes *et al*. 2008). DNA methylation is the most frequently studied epigenetic mark, and is thought to play a key role in the regulation of imprinted genes in mammals (Feil & Berger 2007; Morison *et al*. 2005).

Imprinting during gametogenesis often leads to silencing of the imprinted allele in offspring, and DNA methylation is a key part of this process. DNA methylation is the addition of a methyl group to cytosine residues and normally occurs at CpG (CG dinucleotide on the same strand, with a phosphodiester bond between them, 'p') positions in mammals, but occurs at other sequence motifs in other organisms (Goll & Bestor 2005; Richards 2008). Methylated cytosines are found in

promoter regions, influencing enhancer/promoter interactions as well as within genes themselves (both introns and exons), often interacting with histones to form heterochromatin and regulate gene expression (Cedar & Bergman 2009; Ohlsson *et al*. 2001). CpG islands are DNA sequence regions with a high-density of CpG motifs, that are often differentially methylated to function in the regulation of gene expression. CpG islands were first discovered more than 25 years ago (Bird *et al*. 1985; Tykocinski & Max 1984) and since then it has been shown that 70% of human genes are linked to promoter CpG islands, and that the majority of independent CpG dinucleotides across the human genome are methylated, whereas those associated with CpG islands are frequently not (Weber *et al*. 2007). CpG islands found in the promoter region of genes have been linked to gene expression regulation and can be differentially methylated, or imprinted (Illingworth *et al*. 2010). Such imprinted genes frequently cluster together in genomic regions, or domains, and are controlled by differentially methylated regions (DMRs), or imprinting control regions (ICRs), which can differ in methylation status and be stably methylated over multiple generations (Becker *et al*. 2011; Choufani *et al*. 2011). Variably methylated regions (VMRs) also occur that might be a source of variation for the production of DMRs (Feinberg & Irizarry 2010). The effect of methylation on gene expression depends on many aspects, such as the default state of the DMR, the genomic location of the gene and interactions with other epigenetic mechanisms such as histones and small RNAs.

The *Insulin-like Growth Factor II* (*Igf2*) was one of the first genes to be identified as imprinted in mammmals (Barlow *et al*. 1991; Dechiara *et al*. 1991). *IGF2* is a growth factor that acts through signaling pathways to regulate growth, development and metabolism, causing increased zygotic growth when overexpressed (Vrana 2007; Wood *et al*. 2005). Studies in mice show that the deactivation of *Igf2*

through mutation creates mutants with 60% of normal body weight for paternally transmitted alleles, and phenotypically normal offspring for mutant maternal alleles (DeChiara 1990; 1991). In nature, allele-specific patterns in body size are seen in heterospecific crosses of deer mice, *Peromyscus maniculatus* and it's sister species *P. polionotus*, linked to differential imprinting of several genes (Vrana 2007). In mammals, *IGF2* expression is limited after birth, however in fish expression continues throughout development (O'Neill *et al*. 2007). *Igf2* has several different promoters, depending on the presence of imprinting (Lawton *et al*. 2008), and expresses several different transcripts across development and tissue types (Dechiara *et al*. 1990). The expression of *IGF2* is also regulated by the *mannose-6-phosphate/IGF2* receptor (*Igf2R*), which suppresses offspring growth by encapsulating *IGF2* and transporting it to lysozymes for degradation (Wang *et al*. 1994). Thus *IGF2* is important for the regulation of developmental growth, and the reciprocal imprinting of both *Igf2* and *Igf2R* play a crucial role in this regulation (Killian *et al*. 2001).

The epigenetic regulation of *IGF2* expression has been most intensively studied in mice, where the mechanism has been thoroughly characterised. The *Igf2* domain includes a primary DMR and two secondary DMRs that regulate the expression of *IGF2* along with a long non-coding RNA called *H19*. *Igf2* is located upstream of *H19* separated by ~100 Kb (Murrell *et al*. 2004) with the primary DMR located between them, close to the *H19* promoter. The primary and secondary DMRs interact to produce parent specific 'switches' that regulate gene expression (Murrell *et al*. 2004). The primary DMR binds the CTCF transcription factor, an enhancer that is blocked when the primary DMR is methylated (Bell & Felsenfeld 2000). Methylation of the primary DMR for paternally transmitted alleles, prevents CTCF attachment and

H19 transcription, and allows *Igf2* access to enhancer elements located downstream of *H19*. The primary DMR interacts with a secondary DMR located in exon 6 of *Igf2*. This interaction allows for the expression of *Igf2*, with *H19* being silenced. The maternal allele, however, is unmethylated at the primary DMR and thus bound by CTCF. The primary DMR interacts with another secondary DMR located upstream of the *Igf2* promoter region, creating an inactive chromatin domain in which *Igf2* is located. Thus, *H19* is expressed while *Igf2* is silenced in the maternal allele (Murrell *et al*. 2004).

Imprinting has received less attention in fish than mammals. However, recent work provides evidence for selection on *Igf2* coincident with the evolution of the placenta in matrotrophic fish (O'Neill *et al*. 2007). This study suggested a role for sexual conflict in driving *Igf2* evolution in fish, with expression potentially regulated by genomic imprinting, similar to that in mammals. However, O'Neill *et al* (2007) did not consider duplicate copies of *Igf2* genes that were retained following genome duplication events (Taylor *et al*. 2003). These are two copies of *Igf2* (*Igf2a* and *Igf2b*; Sang *et al*. 2008; Zou *et al*. 2009), which may have subsequently evolved differing functions. This, along with the fact *Igf2* is expressed throughout fish development means that sexual conflict in fish may have led to different mechanisms of *Igf2* expression than in mammals.

The freshwater Goodeid fish are a diverse group of species that demonstrate variable levels of sexual dimorphism (see Appendix I; Macías Garcia *et al*. 2012). This group is thought to have radiated rapidly, around 15 million years ago (Doadrio & Dominguez 2004; Webb *et al*. 2004) and evidence suggests that sexual selection has lead to sexual dimorphism influencing levels of gene flow across species (Ritchie *et al*. 2007; Ritchie *et al*. 2005). *Girardinichthys multiradiatus*, or the Amarillo fish,

is the most sexually dimorphic of the Goodeid fish, indicating that sexual selection might be important in this species. *G. multiradiatus* demonstrates genetic divergence between populations as well as interpopulation variation in sexual dimorphism and male display traits (colourful and enlarged median fins; Macías Garcia *et al*. 2012). Population level crosses between the most geographically divergent populations demonstrate significantly larger offspring size (body weight, width and length) for interpopulation crosses than intrapopulation crosses, with interpopulation size depending on the population-of-origin of each sex (Salvidar *et al*, unpub. results). This suggests that these divergent populations may have evolved differently in response to selection on parental investment in offspring size. This result is not likely to be simply due to hybrid vigour, as this would likely result in an equal size of both interpopulation crosses. Thus, there is the potential for a role of population-specific expression of the *Igf2* gene, regulated by genomic imprinting, in the evolution of *G. multiradiatus* populations.

The aim of this study was to examine the *Igf2* domain in the *G. multiradiatus* system for evidence of population specific patterns of DNA methylation, linked to sexual conflict. To do so required genetic sequence information from the *Igf2* gene. Therefore, the first aim was to obtain the sequence of the *Igf2* copy analysed by O'Neill *et al*. (2007) (that only included the mRNA sequence), including exons and introns, using primers designed from a closely related species, *Ilydon furcidens amacae* (identified as *I. amecae* in O'Neill *et al*. 2007, however this is not officially a species but a subspecies; Kingston 1979). The mechanism of *Igf2* expression is unknown in fish. An orthologous non-coding RNA to the *H19* gene has not been annotated in the fish genomes for which the full sequence information is available (Ensembl Zebrafish: http://www.ensembl.org, search for annotated *H19* gene,

database accessed 21/5/12; Flicek *et al*. 2011), and a BLASTn search of the *H19* sequence from *Mus musculus* against the *Danio rerio* genome did not produce any significantly similar sequences (mouse H19 sequence obtained from the NCBI database: http://www.ncbi.nlm.nih.gov, and a BLASTn search performed against taxid: 7955; Altschul *et al*. 1997). Thus a second aim was to examine sequences for evidence of regulatory features, particularly DMRs (i.e. CpG islands). The final aim was to survey CpG islands for methylation and produce a protocol allowing a quantitative assessment of methylation variation between experimental groups, such as population-of-origin, sex, generation or gamete/soma tissues. The long-term goal is to obtain the entire *Igf2* region, for both duplicate copies, in order to examine allele-specific expression across divergent populations.

METHODS

Design of EPIC primers

To obtain the sequence of *Igf2* in *G. multiradiatus*, Exon Priming Intron Crossing (EPIC) primers (Palumbi & Baker 1994) were designed using a closely related species, *Ilydon furcidens amecae*. The mRNA sequence of the *I. f. amacae Igf2* gene was obtained from O'Neill (2007), and is most likely to be *Igf2b* (see Results). Amplification of the *Igf2* gene was attempted with primers designed using the zebra fish genome, however PCR fragments did not amplify, most likely due to the level of sequence divergence between species. Therefore, primers were designed with the *I. f. amacae Igf2b* mRNA sequence, using Primer3+ (Rozen & Skaletsky 2000). The *Igf2b* mRNA is comprised of four exons. Three sets of primers were designed with one primer in each exon flanking the target intron, thus amplifying each intron as well as

some exon sequence (Table 5.1; Fig. 5.1). Unfortunately all intron 2 amplifications failed (see Results).

Table 5.1 EPIC primers for amplification of introns of *Igf2*, including both sets of intron 2 primers that failed to amplify.

Target region	Primer position	Primer sequence	T_A
Intron 1	Exon 1 Exon 2	F: GATCCGGACACCACTCACTT R: CTCCCCCACACAACGTCTCT	58°C
Intron 2	Exon 2 Exon 3	F: GATGCGCTGCAGTTTGTCT R: GGGTTTGGCACAGTATTGCT	-
Intron 2 (re-designed)	Exon 2 Exon 3	F: CTCACGCTCTACGTTGTGGA R: GACACGTCCCTCTCGGACT	-
Intron 3	Exon 3 Exon 4	F: AGCTGTGACCTCAACCTGCT R: CTTCTTCTGCCACGTTTCGT	58°C

T_A, annealing temperature; -, denotes failed amplification of target sequence.

DNA extraction and amplification of target sequence

DNA was extracted from one *G. multiradiatus* individual from one population (San Juanico, central Mexico) using cell lysis solution (0.1M EDTA, 0.2M Tris pH 8.5 and 1% Tris). 600µl of cell lysis solution was chilled on ice in a 1.5 ml tube. A fin clip (2x3 mm in size) was homogenized and incubated overnight with 3 µl of Proteinase K. After incubation 3 µl of RNaseA was added to the solution and the sample incubated for one hour. Proteins were precipitated with potassium acetate and DNA with 100% isopropanol. DNA was washed with 70% ethanol and eluted in 20 µl of PCR grade water.

PCR for each primer set was carried out in 50 µl reactions; 5 µL of 10× ammonia buffer, 1 µL of 10mM dNTPs, 1 µL of 50mM $MgCl_2$, 1.2 µL of each primer at 30 pM/µL, 0.3 µL of 5U *Taq*, 9.9 µL of Q-Solution (Bioline), 29.5 µL of water and 1 µL of extracted DNA. The thermoprofile for PCR was; 94°C for 3 minutes followed

by 30 cycles of 94°C for 30 seconds, primer specific T_A °C and 72°C for 30 seconds. Final extension was at 72°C for 5 minutes. PCR products were then cleaned up using the MSB Spin PCRapace clean up kit (Thistle Scientific) and then sequenced following a BigDye reaction on an ABI 3730 sequencer. Both forward and reverse strands were sequenced.

Analysis of sequences and bisulfite primer design

Raw DNA sequences were edited and checked by hand using Geneious (Drummond *et al*. 2011), and the edited forward and reverse sequences for each primer set were aligned using MAFFT (Katoh *et al*. 2002). A BLASTn search (Altschul *et al*. 1997) of the NCBI database (http://www.ncbi.nlm.nih.gov) was performed on each region to confirm the sequence identity. TFSEARCH v1.3 (Akiyama 1995) was used to search for potential transcription factor binding sites in conserved intron regions (see Results). TFSEARCH uses the TRANSFAC database (Heinemeyer *et al*. 1998) to search for putative transcription factor binding sites, producing a similarity score to database entries (similarity score threshold for positive match = 85%). Because TFSEARCH is subject to false positives (unverified matches) a phylogenetic footprinting analysis was performed in ConSite (Sandelin *et al*. 2004), using orthologous sequences identified in several species by the BLAST search. Phylogenetic footprinting identifies conserved transcription factor binding sites between species pairs that are likely to be functionally important. Phylogenetic footprinting was performed on conserved regions of intron 1 and intron 3 of *G. multiradiatus* by a pairwise comparison to two and three orthologous sequences respectively (see Fig. 5.2), comparing the binding sites discovered with TFSEARCH,

and using a conservation cut-off of 80%, window size of 50bp and transcription factor score threshold of 80%.

Because genomic imprinting is associated with DNA methylation, a CpG island search was performed using MethPrimer (Li & Dahiya 2002). MethPrimer identifies CpG islands in a given input sequence and designs bisulfite sequencing primers flanking these islands. Islands are commonly defined as 100-200bp windows that have a GC content over a certain level (commonly >50%) and a large observed/expected CpG ratio (Fang *et al*. 2012). The observed/expected ratio (Obs/Exp) is calculated by,

$$Obs/Exp = \frac{Number\ of\ CGs}{(Number\ of\ Cs \times Number\ of\ Gs)} \times Sequence\ length$$

Settings for the CpG island search were; island size = 100bp, GC% > 0.4 and O/E > 0.6 for intron 1 sequence, and island size = 100bp, GC% > 0.3 and O/E > 0.6 for intron 3, the latter settings were relaxed as no CpG islands were discovered at the default CG% of 0.4. The primers used to amplify CpG island regions are presented in Fig. 5.2. Two primer sets were designed, amplifying ~500bp of CpG islands located in intron 1 and intron 3 of *Igf2* (see Results and Fig. 5.1 for positions).

Bisulfite treatment and sequencing

The Qiagen Epitect Bisulfite conversion kit was used to perform bisulfite sequencing, in order to identify methylated nucleotides in the introns of *Igf2*. Briefly, sodium bisulfite was added to genomic DNA and incubated in a buffered solution. Sodium bisulfite deaminates cytosines into uracil (seen as a T upon sequencing), however methylated cytosines are protected from deamination by their methyl group

(remaining a C in the sequence). Recent studies have demonstrated the presence of 5-hydroxymethylcytosine, thought to be an intermediate during the de-methylation of DNA due to its association with CpG islands (Booth *et al*. 2012; Branco *et al*. 2012). Bisulfite sequencing does not distinguish between methylation and hydroxymethylation, however if hydroxymethylation is an intermediate between cytosine and 5-methylcytosine, it is likely to be a functionally important epigenetic mark. Bisulfite sequencing primers were then used to amplify converted DNA for sequencing. Bisulfite sequencing primers were designed to include non-CpG cytosines, which are expected to be unmethylated and are converted to T by sodium bisulfite, thus ensuring that amplified sequence is only from converted DNA. PCR fragments were designed to be <500bp because sodium bisulfite treatment causes fragmentation of genomic DNA.

Tissue was obtained from one *G. multiradiatus* female (somatic tissue) and her offspring (a developing zygote) and genomic DNA was extracted. The zygote came from an intrapopulation cross from the San Matias el Grande (SMG; see Appendix I) population. Although more extensive interpopulation crosses were planned, unfortunately most of the females did not produce offspring. PCR of bisulfite converted DNA was carried out for mid-intron regions of introns 1 and 3, in which CpG islands were detected (see Results), for both the embryo and somatic tissue of the female.

Unspecific priming and background noise during PCR can yield unreliable sequence traces that do not allow for analysis of directly sequenced PCR products (Jiang *et al*. 2010). Such background noise means that direct sequencing of bisulfite converted PCR product is rarely performed (Myohanen *et al*. 1994; Paul & Clark 1996). Instead, methylation is commonly identified through the cloning of PCR

products. However, this relies on the probability of selecting a representative sample of clones from the sequence population of PCR products, and thus a large number of clones need to be sequenced (>10 clones per sample; Jiang *et al*. 2010). Direct sequencing of PCR products is more quantitative because sequences represent the overall pool of products. Although some novel methods have recently been developed (e.g. Lewin *et al*. 2004), a quick, easy and inexpensive protocol such as that carried out by Jiang (2010) is more desirable for large-scale surveys of methylation. Therefore, an optimised PCR protocol was performed, analogous to the method in Jiang *et al*. (2010), with cloning used to confirm the methylation status of each region and direct sequencing performed for a pilot study (see below). A hot-start *Taq* was used to reduce unspecific priming along with Q-solution, a PCR additive that helps to reduce non-specific amplification of DNA. Each PCR contained 15 μL PCR mix (Qiagen), 0.4 μL of forward and reverse primer, 13.2 μL of PCR grade water, 5 μL of Q-solution and 1 μL of bisulfite converted DNA to make a 35 μL reaction. PCR reactions were checked on agarose gel to confirm correct amplification of fragment lengths, and each product was cloned using a TOPO TA cloning kit (Invitrogen). Ten colonies from each of the four PCR reactions were picked for purification and sequencing. From these clones the methylation status of nucleotide positions were established and were compared to the results of directly sequenced PCR products.

A pilot study was carried out using direct sequencing to examine methylation variation across multiple samples. The main aim of this pilot was to establish the level of technical variation caused by direct sequencing and compare this measure to the cloned sequences. One male and one female from two divergent populations, Zempoala (Z) and San Matías el Grande (SMG) (see Appendix I, Fig. 2 for locations), were surveyed for intron 3, which demonstrated a high level of DNA methylation.

The Zempoala and San Matías Grande populations are geographically separated and demonstrate significant differences in offspring size in crosses, depending on the population-of-origin of the each parent (see Introduction). Technical replication enabled an assessment of the repeatability of the direct bisulfite sequencing protocol. Three replicate muscle tissue samples from each individual were processed independently, from the DNA extraction step through to sequencing. Sequencing was carried out directly from PCR products, quantified using the relative peak heights of Cs and Ts at each position. Percentage methylation was calculated as the peak height of C divided by the heights of (C+T). Peak heights were measured using the 4peaks software (Griekspoor & Groothuis 2006).

The population genetic variation in genomic (i.e. non-converted) sequences of intron 3 was also surveyed in both populations, to associate any genetic variation with epigenetic variation in the sequenced region. Chromatograms were examined and edited by hand using Geneious (Drummond *et al*. 2011) and aligned using MAFFT (Katoh *et al*. 2002).

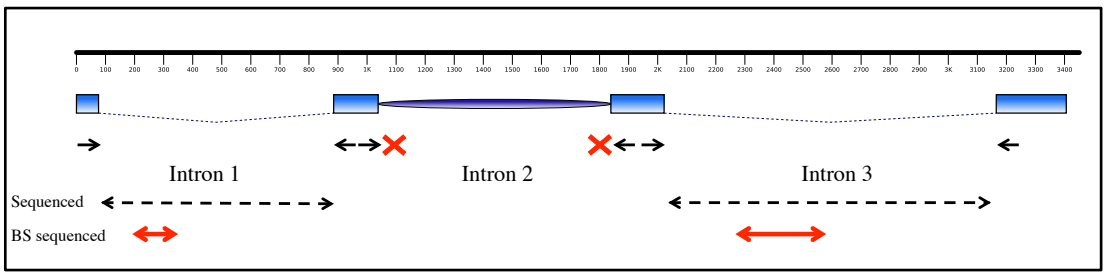


Fig. 5.1 Summary of sequenced regions of the *Igf2* gene in *G. multiradiatus* using EPIC and bisulfite sequencing primers. Small black arrows denote EPIC primer positions. Exons are blue boxes and introns dashed hats. The purple oval represents intron 2, which is of unknown size due to failed amplification. Dashed arrows show successfully sequenced regions and red arrows bisulfite sequenced regions.

RESULTS

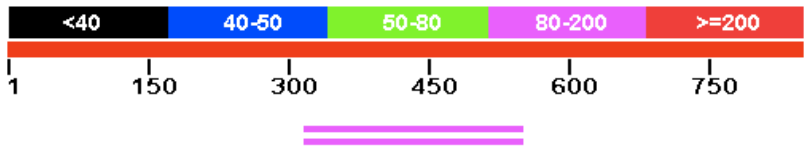
EPIC primers were successful in sequencing introns 1 and 3, and also captured 37bp of exon 2 and 52bp of exon 3 in each sequence respectively. Intron 2 failed to amplify, producing a high molecular weight band when ran on a gel. Primers were re-designed in different regions of exon 2 and exon 3 in an attempt to obtain intron 2, however these also failed to amplify.

Intron 1 of the *Igf2* gene was 923bp in length and intron 3 1272bp long. Pairwise identity of aligned protein sequences of *IGF2* demonstrated a 58.2% and 70.7% identity of *Igf2a* and *Igf2b* respectively, compared to the *Igf2* copy in *I. amacae*. This suggests that the version of *Igf2* examined in O'Neill *et al* (2007), and thus the copy amplified here was most likely *Igf2b*. BLAST searches of each intron sequence revealed highly conserved regions in the center of each intron (Fig. 5.2). Intron 1 showed significant hits to intron sequences in the *Igf2* gene in two different species of cichlid fish. This conserved region was 236bp in length with 81% identity to intron regions in these species (Fig. 5.2). Intron 3 showed a conserved region 206bp in length with significant matches to three species of fish; cichlid, seabass and largemouth bass, with >83% identity to each (Fig. 5.2). None of these BLAST hits indicated which copy of *Igf2* the sequences originated from. However, these results are suggestive of functional conservation of mid-intron sequence motifs, which might play a role in gene expression regulation or alternative splicing.

A transcription factor binding site search was carried out using TFSEARCH. In total, 24 and 18 binding sites were discovered in the conserved regions of intron 1 and intron 3 respectively. The functions of these transcription factors are broad, but generally they are involved in development, cellular growth and differentiation, particularly haematopoietic development. Four specific transcription factor binding

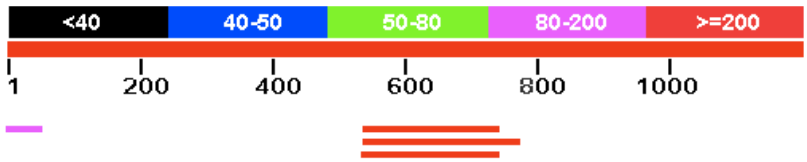
sites were present in both of the intron conserved regions, as well as identified as phylogenetically conserved between all species comparisons using ConSite. These were the binding sites for GATA-1, MZF1, SRY and USF (Table 5.2).

A



Significant BLAST matches	Coverage	E value	Max ident
Oreochromis niloticus Igf2 gene, complete cds	27%	4e-42	81%
Tilapia mossambica Igf2 gene	27%	6e-41	81%

B



Significant BLAST matches	Coverage	E value	Max ident
Oreochromis niloticus Igf2 gene, complete cds	17%	2e-56	86%
Lates calcarifer Igf2 gene, complete cds	19%	1e-49	83%
Micropterus salmoides Igf2 gene, complete cds	17%	1e-48	84%
Ilyodon amecae Igf2 gene, complete cds	4%	4e-14	96%

Fig. 5.2 BLASTn results for (a) intron 1 and (b) intron 3 of *Igf2b* in *G. multiradiatus* (see Methods for details). Positions of significant matches are shown on the chart above with the corresponding species in order in the table below. The small region near the beginning of intron 3 corresponds to exon 3 of *I. f. amecae* from which primers were designed.

CpG island searches were carried out using MethPrimer to identify potentially methylated regions within each intron. MethPrimer identified three and two CpG islands located within each intron and intron/exon boundaries in introns 1 and 3 respectively (Fig. 5.3). Interestingly, putative CpG islands were discovered directly upstream of each mid-intron conserved region, indicating a potential region of methylation coincident with conserved sequence regions. This might indicate

functional sequences with a potential role in gene expression, similar to the DMRs seen in mammalian *Igf2* genes.

A

O. niloticus	*T. mossambica*
GATA-1 (1)	GATA-1 (2)
GATA-2 (1)	GATA-2 (2)
GATA-3 (1)	GATA-3 (2)
MZF1 (2)	MZF1 (2)
SRY (1)	SRY (1)
USF (2)	USF (1)

B

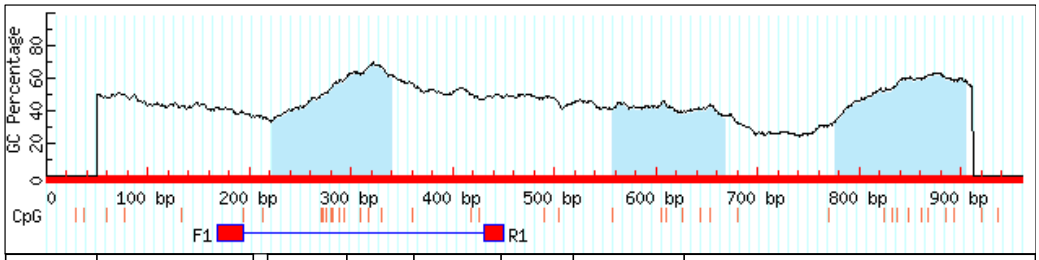
O. niloticus	*L. calcarifer*	*M. salmoides*
GATA-1 (1)	-	-
Nkx (2)	Nkx (1)	Nkx (2)
MZF1 (3)	MZF1 (2)	MZF2 (2)
SRY (2)	SRY (1)	SRY (1)
USF (2)	USF (2)	USF (2)

Table 5.2 ConSite, phylogenetic footprinting results for conserved regions of (a) intron 1 and (b) intron 3, demonstrating conserved transcription factor binding sites between *G. multiradiatus* and each species shown. Each species in the analysis was identified through the BLAST search of the NCBI database. Numbers in brackets are the number of occurrences of that transcription factor binding site within each conserved intron region. See Supplementary Table S5.1 for a summary of the function of each transcription factor.

MethPrimer was used to design primers to amplify bisulfite converted DNA from the CpG islands upstream of the conserved intron regions (Fig. 5.3). Cloned sequences of Intron 1 demonstrated no DNA methylation in the predicted CpG island, however methylation was confirmed within intron 3, with methylation occurring at every CpG position within the predicted CpG island (Fig. 5.4). Seven methylated CpG dinucleotides were discovered through cloning and sequencing of the intron 3 PCR product. Variation in the level of methylation was seen between CpG positions, ranging from complete methylation (100%), to 60% at position '259bp' (Fig. 5.4). This position demonstrated a pattern of methylation expected if only one allele was methylated and thus is a possible candidate for allele-specific methylation. These

sequences were obtained from one female from the SMG population and her offspring, and little difference in methylation was seen between them. Direct sequencing of PCR product was undertaken to examine methylation in a more quantitative manner.

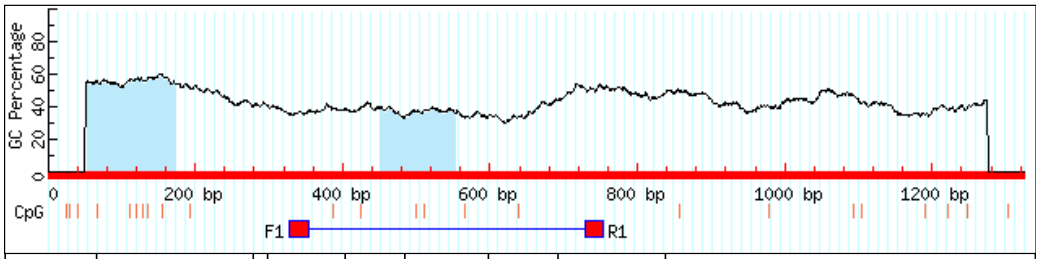
A



	Size (start-end)
Island 1	118bp (223-340)
Island 2	111bp (558-668)
Island 3	129bp (776-904)

Primer	Size	T_m (°C)	GC%	Num. 'C's	Sequence
F1	25	55	44	4	AGAATGTATAAATTGTAAAGGGAGA
R1	20	57	55	4	CCACCCCAAAAAATTACAAA
Product size: 281bp, CpGs in product: 15					

B



	Size (start-end)
Island 1	121bp (54-174)
Island 2	102bp (452-553)

Primer	Size	T_m (°C)	GC%	Num. 'C's	Sequence
F1	27	60	44	4	TGTTGATTATGGTGGAAAATATTAGTG
R1	25	59	52	4	CAATACAACCTACAACCATCAACTT
Product size: 418bp , CpGs in product: 6					

Fig. 5.3 CpG island search results and design of bisulfite sequencing primers using MethPrimer (Li & Dahiya 2002), for the entirety of (a) intron 1 and (b) intron 3 sequences. Graphs show the intron regions in which primers were designed. Blue shading indicates CpG islands, red ticks indicate CpG positions and the selected primers are shown below each graph as red boxes. Below each chart are tables summarizing the CpG island details and the parameters for the selected bisulfite primers. Primers for the 3' CpG islands of intron 1 and 5' CpG island of intron 3 were designed, however this study focuses on the CpG islands coincident with the conserved mid-intron regions.

Direct sequencing of PCR products produced clean sequence traces with little technical variation. Figure 5.5 presents the results of the direct sequencing of an SMG

female compared to the cloned sequences (Fig. 5.5a) and levels of methylation between directly sequenced samples, including the level of technical error around the mean (Fig. 5.5b). The direct sequencing results closely resembled those of the cloned PCR products (Fig. 5.5a). Raw chromatograms from directly sequenced samples were of high quality and background noise was minimal, allowing for accurate quantification of peak heights at CpG positions. Two positions were not included in the direct sequencing results (the first and last CpGs in the cloned sequences) due to trimming of low quality sequence near the primer sites. Low levels of technical variation were seen between directly sequenced technical replicates (Fig. 5.5b), suggesting that direct sequencing can be used to quantify methylation between experimental treatments. Interestingly these results indicate potential population specific levels of methylation at several positions, however this patterns needs to be verified with increased biological replication.

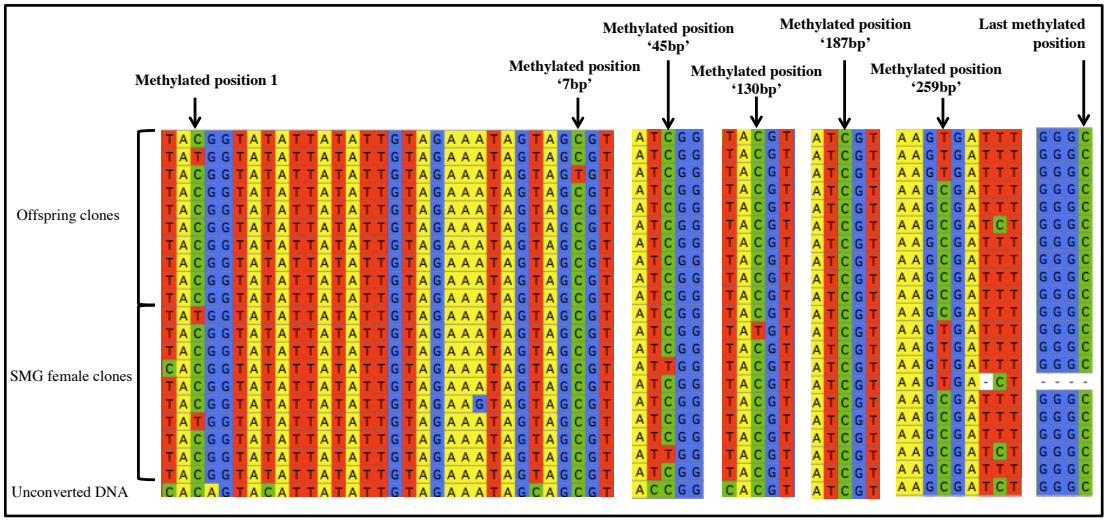


Fig. 5.4 Multiple alignment of bisulfite treated and cloned DNA from intron 3 of *Igf2b*, showing each methylated CpG position (white space within alignment indicates intervening sequence between CpG positions). The first ten rows are clones from offspring tissue and the second ten sequences from the mother, an SMG female. The last row is unconverted (genomic) DNA. Each methylated CpG is indicated with an arrow.

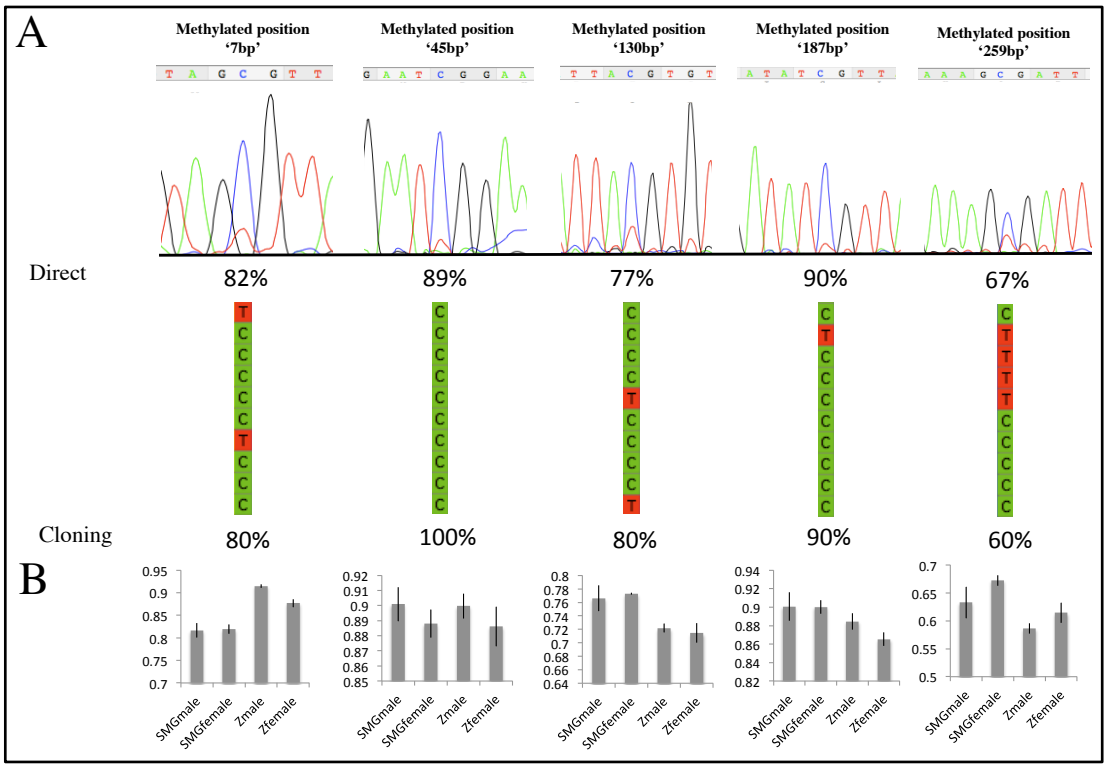


Fig. 5.5 Comparison of direct sequencing (raw traces are presented with percentage methylation below each position) and cloning methods (columns demonstrate the number of methylated clones for each position, with percentage methylation below each one) for CpG detection (a), and examination of technical variation associated with direct sequencing for methylation quantification (b). Bar plots show the mean and standard error of percentage methylation for three technical replicates per sample. Only the 5 methylated CpGs captured by direct sequencing are shown.

Direct sequencing uncovered one position that appeared to be a differentially methylated between populations. This position was methylated in the Zempoala population (occurring just upstream of position '130bp') but not in the SMG group, due to a single nucleotide polymorphism (SNP) that changed the CpG to a CpA. To examine this on a genetic level, several individuals from each population were sequenced for intron 3 and aligned with the bisulfite-converted sequences (Fig. 5.6). This survey uncovered genetic variation at the polymorphic position in both populations for the G nucleotide, which allows CpG methylation, with this nucleotide

being at a low frequency. Thus the seemingly population-specific methylation seen in the bisulfite-converted sequences was in fact due to sampling bias. The level of genetic variation between genomic sequences was low and associated with CpG positions. Only four sites were variable in the 750bp region, with two of these associated with a CpG motif. Interestingly, both these sites were associated with the CpG motif through SNPs at the G positions, rather than the cytosine. One of these CpGs was the position reported above, and the other was the first CpG position in the CpG island, also a G/A polymorphism, which showed a higher frequency of G nucleotides than A in both populations. This indicates that there is genetic variation for CpG dinucleotides that allow for the creation of methylation marks. Further, there is genetic variation for these two methylated positions in both populations that seems to be maintained at similar frequencies in each. Figure 5.7 summarizes the results found in intron 3 of *Igf2b* of *G. multiradiatus*.

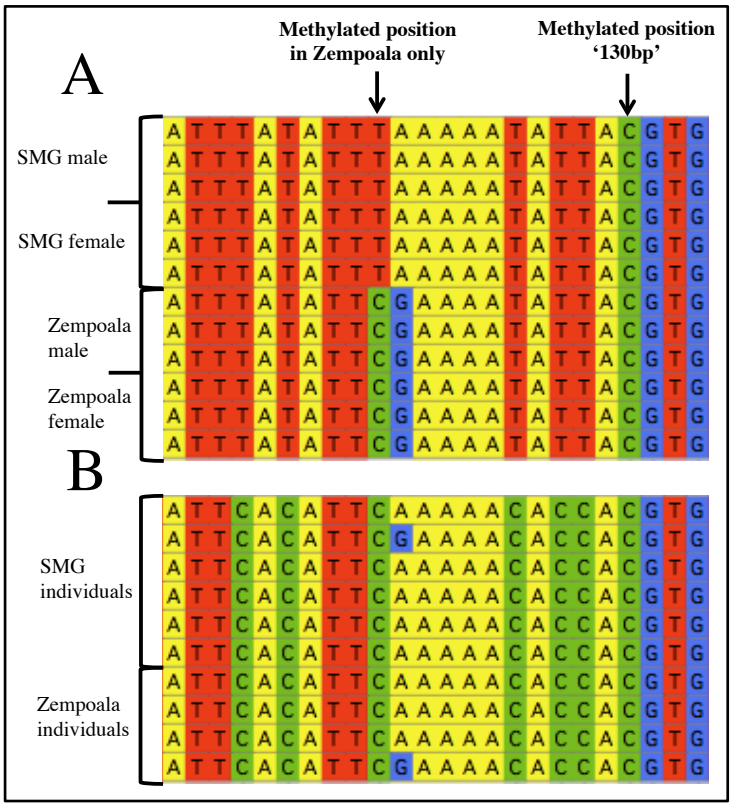


Fig. 5.6 Alignment of bisulfite treated (a) and unconverted genomic DNA (b) for San Matías Grande (SMG) and Zempoala populations, showing that both genetic and epigenetic variation occur within populations. Low frequency SNPs occur in both populations, with CpG dinucleotides being methylated and CpAs being unmethylated.

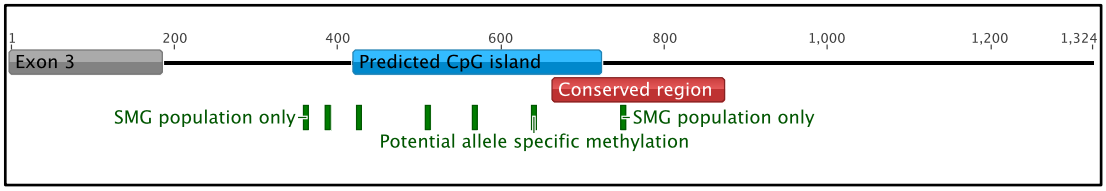


Fig. 5.7 Summary of results for intron 3 of the *Igf2b* gene in *G. multiradiatus*. Each green bar represents a methylated CpG dinucleotide, the red bar is the mid-intron conserved region and blue bar the predicted CpG island.

DISCUSSION

Sexual conflict is an antagonistic evolutionary process thought to result in parent-of-origin effects that are often controlled through genomic imprinting (Murphy & Jirtle 2003). Differentially methylated regions control the expression of the *Igf2* gene in mammals, and strong signals of selection have recently been detected in *Igf2* in teleost fish, most likely driven by sexual conflict (O'Neill *et al*. 2007). Epigenetic variation is associated with imprinting and is also known to be an important regulator of gene expression. Here, sequences were obtained from *Igf2b* in *G. multiradiatus*, and DNA methylation was discovered in intron 3 that resembled DMRs previously discovered in mammals. Coincident with these methylation patterns were structurally conserved sequence regions, indicating potential gene regulatory motifs. Methylation patterns varied between samples, with very little technical variation, allowing for the future examination of quantitative differences in methylation by direct sequencing of PCR products.

Sequencing of the Igf2b *gene*

The first aim of this study was to obtain sequence information for the *Igf2b* gene in the Amarillo fish, *G. multiradiatus*. The use of EPIC primers allowed successful amplification of two out of three introns and some of the flanking exons. However, because intron 2 did not successfully amplify it was not possible to design primers and sequence back across the exons, to obtain full exon sequences. The failure of intron 2 to amplify might have been due to sequence divergence in the primer attachment sites. However, sequence divergence of *Igf2b* between *I. amacae*, in which primers were designed, and *G. multiradiatus* for the partial exon sequences obtained was low. It is possible that the 3' end of exon 2 and the 5' of exon 3, which were not

obtained, may demonstrate higher levels of divergence. However, it is more likely that intron 2 is simply very large and thus difficult to obtain with conventional PCR. Even after accounting for a large fragment size (by increasing cycling times) this region did not amplify. The second intron of *Igf2b* in *D. rerio* is 1,881bp long, indicating that the *G. multiradiatus Igf2b* second intron might also be large. Although conventional *Taq* polymerase can process fragments of up to 5Kb the efficiency of the reaction tends to reduce with longer fragments, and thus intron 2 might be obtained if a polymerase appropriate for longer fragment lengths is used.

Intronic DNA methylation

Evidence of DNA methylation in intron 3 of *Igf2b* in *G. multiradiatus* suggests the potential for epigenetic regulation of this gene, analogous to the mechanisms known in mammals. The lack of an annotated *H19* in fish species for which full genomes are available and the conservation of *H19* across all mammals (Smits *et al*. 2008; Stadler 2010) suggests that the mechanism of expression of *Igf2b* in fish might differ from that in mammals. Methylation patterns within introns have previously been associated with the regulation of alternative splicing, through interactions with histones (Cedar & Bergman 2009). Intronic splicing regulatory elements (ISREs) can influence the inclusion or exclusion of exons, and are crucial for context dependent alternative splicing (Yeo *et al*. 2007). Thus the patterns of DNA methylation seen here, associated with potential regulatory elements, might be involved in alternative splicing regulation. Although there is evidence for methylation control at exonic splicing enhancers, there is little data on the association of methylation patterns with ISREs.

Several transcription factor binding sites were discovered in the conserved mid-intron region of intron 1 and intron 3, that were conserved between several fish species. These binding sites were associated with several ubiquitously expressed transcription factors that function in cellular growth and differentiation, playing multiple regulatory roles during development. For example, SRY (Sex Determining Region Y) is a conserved transcription factor found on the Y chromosome in mammals that is responsible for sex determination, through initiating the formation of testes in males (Jazin & Cahill 2010). GATA-1, MZF1, SRY and USF binding sites are commonly found in the promoter regions of genes (Coulibaly *et al*. 2006; Yan *et al*. 2005), and have also been shown to reside in the introns of several genes, including *FBN1* in primates (Jackson-Hayes *et al*. 2003; Singh *et al*. 2008). Most transcription factor binding sites associated with influencing gene expression tend to occur in the first or second intron of genes (Rose, 2008), thus it is interesting that intron 3 of *Igf2b* has a putative regulatory region that is associated with a methylated CpG island.

USF is a transcription factor that commonly regulates gene expression through promoter binding, and there is evidence that USF is prevented from attaching to binding sites when DNA methylation is present (Hou et al, 2012; Huji et al, 2006). Further, transcription factors such as GATA-1 play a direct role in histone modifications and chromatin regulation, leading to gene expression repression or activation (Letting *et al*. 2003). Therefore, it is possible that the conserved region within intron 3 (and potentially intron 1) is a regulatory feature that interacts with DNA methylation to influence the expression of *Igf2b* (Tate & Bird 1993). The evolution of large introns that harbour regulatory elements might be common in fish, with a recent study demonstrating a greater number of large introns (500-2000bp) in

D. rerio than other teleost fish, caused by an ancient intron size expansion (Moss *et al*., 2011).

Another potential (or parallel) role for the methylated conserved region in intron 3 might be as a DMR influencing parent-specific expression of *Igf2b*. DMRs in mice are located within the *Igf2* gene and are important for imprinting controlled gene expression (Murrell *et al*. 2004). Thus the methylation in intron 3 detected here, coincident with a conserved sequence element, might be a DMR involved in expression regulation of *Igf2b* in fish. The sequenced copy of *Igf2* demonstrated similarity to copy *b* and this duplicate gene is thought to have a more limited expression pattern than *Igf2a*. *IGF2b* expression has been detected during embryogenesis in zebra fish but adult expression seems limited to the liver tissues, although both copies of *Igf2* show biological effects when injected into zebrafish embryos (Zou *et al*. 2009). The retention of duplicate copies of *Igf2* suggests functionality of each gene, and thus it will be necessary to investigate both *Igf2* copies in future studies. This will include obtaining the full *Igf2* domain of each copy, bisulfite sequencing of each to identify DNA methylation patterns, uncovering other putative DMRs, and quantification of their expression levels.

The direct sequencing protocol showed comparable results to cloned sequences, with little technical variation, allowing a quantitative examination of methylation patterns. Technical variation can occur due to incomplete bisulfite conversion or artifacts introduced during the amplification of target fragments, and often the use of fragmented DNA due to bisulfite conversion necessitates a second round of PCR using nested primers. Methylation was discovered spanning the CpG island directly upstream from the conserved sequence region in intron 3, through cloning of PCR products. Direct sequencing confirmed these results, providing a

quicker and cheaper method than cloning for examining methylation patterns to base-pair resolution. The optimized protocol for direct sequencing of bisulfite converted PCR products demonstrated here allows large numbers of samples to be processed. Further, the use of well-designed primers and a hot-start *Taq* meant that only one round of PCR was necessary to obtain PCR product for sequencing. Therefore this protocol is repeatable and accurate for methylation quantification and a full study with multiple treatment groups is likely to produce biologically meaningful data that can be analysed statistically.

Interestingly, position '259bp' was located immediately upstream of the conserved region of intron 3, indicating a role for this region in imprinted regulation of gene expression. Position '259bp' is potentially differentially imprinted in male and female gametes and thus may be a result of sexual conflict. The location of this methylation suggests it might be involved in regulating the conserved sequence region directly downstream of the CpG island. However, further work is needed to characterize this region and ascertain what proteins are binding to the sequence in order to infer its function, and to discover how it interacts with the CpG island. Phenotypic variation in body size across *G. multiradiatus* populations is suggestive of population-specific mode of *IGF2* regulation. If patterns of methylation across the entire *Igf2b* domain could be obtained for both the Zempoala and San Matías Grande populations, this might uncover the population specific regulation of *Igf2b*.

Imprinting in the Igf2 *gene*

Future work on the *Igf2* domain in *G. multiradiatus* should include (i) the full characterization of imprinting regulated gene expression and (ii) experiments to examine sexual conflict between populations. Characterizing the mode of *Igf2*

expression would provide details that might generalize to other teleost fish and uncover the differing roles of each *Igf2* duplicate. Here, short lengths of exon sequence were obtained. These sequences can now be used to design primers in order to amplify the full mRNA transcript of *Igf2b* using Rapid Amplification of cDNA Ends (RACE) PCR. This will provide exon sequence information that will allow for the addressing of the two main questions above. Sequences can be used to design degenerate primers in order to obtain *Igf2a* from *G. multiradiatus*. Using sequence information from both copies of *Igf2*, primers for quantitative PCR and methylation variation can be designed, to full characterise the mechanism of expression of each duplicated gene.

One key prediction for imprinting in sexual conflict is that offspring demonstrate methylation patterns according to the population of origin of each sex (assuming population-specific patterns of imprinting). Another key prediction for epigenetic mediated sexual conflict is the differential methylation of conflict genes between gametes (see Introduction). Thus to examine methylation patterns in relation to sexual conflict it is necessary to examine methylation patterns in parental gametes and zygotic offspring (*Igf2* is most highly expressed during zygote development, influencing adult body size). Thus future work should examine both intra- and interpopulation crosses of the SMG and Z populations, tracking parental (gametic) patterns into the offspring in order to observe the population specific effects in interpopulation crosses. All of the four possible population crosses should be performed (within and between SMG and Z), along with replicate families of each cross for biological replication. Using direct sequencing, multiple replicates of bisulfite treated samples can be cheaply and quickly obtained. Thus, methylation and gene expression patterns of both copies of the *Igf2* gene could be examined in parental

gametes and their offspring, for each of the four possible mating combinations, and correlated with phenotypic data on body size. Biological replicates would allow for the statistical analysis of the expression data, methylation patterns and phenotypic data. Thus multiple types of complementary data can be jointly analysed, to uncover the role of sexual conflict in the divergence of *G. multiradiatus* populations.

Investigating patterns of methylation within and between populations will help to uncover mechanisms of gene expression, but might also aid in the investigation of population epigenetic variation (Richards 2008). Here, the discovery of genetic variation for CpG dinucleotides indicates that genetic variation and epigenetic variation might interact. Thus, selection may act on genetic and epigenetic variation in populations as has been recently suggested (Feinberg & Irizarry 2010). This has consequences for the evolution of non-coding DNA, and several recent studies have shown that some introns demonstrate signals of natural selection, particularly purifying selection, in splice sites and regulatory motifs (Andolfatto 2005; Farlow *et al*. 2012; Gazave *et al*. 2007; Zhu *et al*. 2009). The study of epigenetics in evolution is still in the early stages, however an increasing number of studies are now examining epigenetic variation in order to elucidate its role in evolutionary processes.

REFERENCES

Akiyama Y (1995) TFSEARCH: Searching transcription factor binding sites. http://www.rwcp.or.jp/papia.

Altschul SF, Madden TL, Schaffer AA, *et al*. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*, **437**, 1149-1152.

Barlow DP, Stoger R, Herrmann BG, Saito K, Schweifer N (1991) The mouse insulin-like growth-factor type-2 receptor is imprinted and closely linked to the *Tme* Locus. *Nature*, **349**, 84-87.

Bartolomei MS, Tilghman SM (1997) Genomic imprinting in mammals. *Annual Review of Genetics*, **31**, 493-525.

Becker C, Hagmann J, Muller J, *et al*. (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature*, **480**, 245-U127.

Bell AC, Felsenfeld G (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature*, **405**, 482-485.

Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A (2009) An operational definition of epigenetics. *Genes & Development*, **23**, 781-783.

Bird A (2007) Perceptions of epigenetics. *Nature*, **447**, 396-398.

Bird A, Taggart M, Frommer M, Miller OJ, Macleod D (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, **40**, 91-99.

Bonduriansky R, Chenoweth SF (2009) Intralocus sexual conflict. *Trends in Ecology & Evolution*, **24**, 280-288.

Booth MJ, Branco MR, Ficz G, *et al*. (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, **336**, 934-937.

Branco MR, Ficz G, Reik W (2012) Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature Reviews Genetics*, **13**, 7-13.

Brandvain Y, Van Cleve J, Ubeda F, Wilkins JF (2011) Demography, kinship, and the evolving theory of genomic imprinting. *Trends in Genetics*, **27**, 251-257.

Cedar H, Bergman Y (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, **10**, 295-304.

Chapman T (2006) Evolutionary conflicts of interest between males and females. *Current Biology*, **16**, R744-R754.

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

Choufani S, Shapiro JS, Susiarjo M, *et al*. (2011) A novel approach identifies new differentially methylated regions (DMRs) associated with imprinted genes. *Genome Research*, **21**, 465-476.

Coulibaly I, Gahr SA, Palti Y, Yao J, Rexroad CE, 3rd (2006) Genomic structure and expression of uncoupling protein 2 genes in rainbow trout (*Oncorhynchus mykiss)*. *BMC Genomics*, **7**, 203-216.

Coyne JA, Orr HA (2004) *Speciation* Sinauer, Sunderland, Mass.

Dechiara TM, Efstratiadis A, Robertson EJ (1990) A growth-deficiency phenotype in heterozygous mice carrying an *insulin-like growth factor-II* gene disrupted by targeting. *Nature*, **345**, 78-80.

Dechiara TM, Robertson EJ, Efstratiadis A (1991) Parental imprinting of the mouse *insulin-like growth factor II* gene. *Cell*, **64**, 849-859.

Doadrio I, Dominguez O (2004) Phylogenetic relationships within the fish family Goodeidae based on cytochrome b sequence data. *Molecular Phylogenetics and Evolution*, **31**, 416-430.

Drummond AJ, Ashton B, Buxton S, *et al*. (2011) *Geneious v5. 4*. Biomatter Ltd Auckland, New Zealand.

Fang WJ, Zheng Y, Wu LM, *et al*. (2012) Genome-wide analysis of aberrant DNA methylation for identification of potential biomarkers in colorectal cancer patients. *Asian Pacific Journal of Cancer Prevention*, **13**, 1917-1921.

Farlow A, Dolezal M, Hua LS, Schlotterer C (2012) The genomic signature of splicing-coupled selection differs between long and short introns. *Molecular Biology and Evolution*, **29**, 21-24.

Feil R, Berger F (2007) Convergent evolution of genomic imprinting in plants and mammals. *Trends in Genetics*, **23**, 192-199.

Feinberg AP, Irizarry RA (2010) Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences, USA*, **107**, 1757-1764.

Flicek P, Amode MR, Barrell D, *et al*. (2011) Ensembl 2011. *Nucleic Acids Research*, **39**, D800-D806.

Funk DJ, Nosil P, Etges WJ (2006) Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. *Proceedings of the National Academy of Sciences, USA*, **103**, 3209-3213.

Garnier O, Laoueille-Duprat S, Spillane C (2008) Genomic imprinting in plants. *Epigenetics*, **3**, 14-20.

Gavrilets S, Hayashi TI (2005) Speciation and sexual conflict. *Evolutionary Ecology*, **19**, 167-198.

Gavrilets S, Waxman D (2002) Sympatric speciation by sexual conflict. *Proceedings of the National Academy of Sciences, USA*, **99**, 10533-10538.

Gazave E, Marques-Bonet T, Fernando O, Charlesworth B, Navarro A (2007) Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biology*, **8**, R21-R21.13.

Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*, **74**, 481-514.

Gregg C, Zhang JW, Butler JE, Haig D, Dulac C (2010a) Sex-specific parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 682-685.

Gregg C, Zhang JW, Weissbourd B, *et al*. (2010b) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 643-648.

Griekspoor A, Groothuis T (2006) 4Peaks version 1.7.2.

Haig D (2000) The kinship theory of genomic imprinting. *Annual Review of Ecology and Systematics*, **31**, 9-32.

Haig D, Westoby M (1989) Parent-specific gene-expression and the triploid endosperm. *American Naturalist*, **134**, 147-155.

Heinemeyer T, Wingender E, Reuter I, *et al*. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Research*, **26**, 362-367.

Illingworth RS, Gruenewald-Schneider U, Webb S, *et al*. (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genetics*, **6**, 1-15.

Jackson-Hayes L, Song S, Lavrentyev EN, *et al*. (2003) A thyroid hormone response unit formed between the promoter and first intron of the carnitine

palmitoyltransferase-I∝ gene mediates the liver-specific induction by thyroid hormone. *Journal of Biological Chemistry*, **278**, 7964-7972.

Jazin E, Cahill L (2010) Sex differences in molecular neuroscience: from fruit flies to humans. *Nature Reviews Neuroscience*, **11**, 9-17.

Jiang MH, Zhang YH, Fei J, *et al*. (2010) Rapid quantification of DNA methylation by measuring relative peak heights in direct bisulfite-PCR sequencing traces. *Laboratory Investigation*, **90**, 282-290.

Johannes F, Colot V, Jansen RC (2008) Epigenome dynamics: a quantitative genetics perspective. *Nature Reviews Genetics*, **9**, 883-890.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059-3066.

Killian JK, Nolan CM, Stewart N, *et al*. (2001) Monotreme *IGF2* expression and ancestral origin of genomic imprinting. *Journal Of Experimental Zoology*, **291**, 205-212.

Kingston D (1979) *Behavioral and morphological studies of the goodeid genus Ilyodon, and comparative behavior of fishes of the family Goodeidae*, University of Michigan.

Kirkpatrick M, Nuismer SL (2004) Sexual selection can constrain sympatric speciation. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **271**, 687-693.

Lawton BR, Carone BR, Obergfell CJ, *et al*. (2008) Genomic imprinting of *IGF2* in marsupials is methylation dependent. *BMC Genomics*, **9**, 1-11.

Letting DL, Rakowski C, Weiss MJ, Blobel GA (2003) Formation of a tissue-specific histone acetylation pattern by the hematopoietic transcription factor GATA-1. *Molecular and Cellular Biology*, **23**, 1334-1340.

Lewin J, Schmitt AO, Adorjan P, Hildmann T, Piepenbrock C (2004) Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics*, **20**, 3005-3012.

Li LC, Dahiya R (2002) MethPrimer: designing primers for methylation PCRs. *Bioinformatics*, **18**, 1427-1431.

Lloyd VK, Sinclair DA, Grigliatti TA (1999) Genomic imprinting and position-effect variegation in *Drosophila melanogaster*. *Genetics*, **151**, 1503-1516.

Maan ME, Seehausen O (2011) Ecology, sexual selection and speciation. *Ecology Letters*, **14**, 591-602.

Macías Garcia C, Ramirez E (2005) Evidence that sensory traps can evolve into honest signals. *Nature*, **434**, 501-505.

Macías Garcia C, Smith G, Zuarth CG, Graves JA, Ritchie MG (2012) Variation in sexual dimorphism and assortative mating do not predict genetic divergence in the sexually dimorphic Goodeid fish *Girardinichthys multiradiatus*. *Current Zoology*, **58**, 440-452.

Martin CC, Mcgowan R (1995) Parent-of-origin specific effects on the methylation of a transgene in the zebrafish, *Danio rerio*. *Developmental Genetics*, **17**, 233-239.

Martin OY, Hosken DJ (2003) The evolution of reproductive isolation through sexual conflict. *Nature*, **423**, 979-982.

M'Gonigle LK, Mazzucco R, Otto SP, Dieckmann U (2012) Sexual selection enables long-term coexistence despite ecological equivalence. *Nature*, **484**, 506-509.

Morison IM, Ramsay JP, Spencer HG (2005) A census of mammalian imprinting. *Trends in Genetics*, **21**, 457-465.

Moss SP, Joyce DA, Humphries S, Tindall KJ, Lunt DH (2011) Comparitive analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage. *Genome Biology and Evolution*. **3**, 1187-1196.

Murphy SK, Jirtle RL (2003) Imprinting evolution and the price of silence. *Bioessays*, **25**, 577-588.

Murrell A, Heeson S, Reik W (2004) Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nature Genetics*, **36**, 889-893.

Myohanen S, Wahlfors J, Janne J (1994) Automated fluorescent genomic sequencing as applied to the methylation analysis of the human ornithine decarboxylase Gene. *DNA Sequence*, **5**, 1-8.

O'Neill MJ, Lawton BR, Mateos M, *et al*. (2007) Ancient and continuing Darwinian selection on *insulin-like growth factor II* in placental fishes. *Proceedings of the National Academy of Sciences, USA*, **104**, 12404-12409.

Ohlsson R, Renkawitz R, Lobanenkov V (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in Genetics*, **17**, 520-527.

Palumbi SR, Baker CS (1994) Contrasting population-structure from nuclear intron sequences and mtDNA of humpback whales. *Molecular Biology and Evolution*, **11**, 426-435.

Parker GA, Partridge L (1998) Sexual conflict and speciation. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **353**, 261-274.

Paul CL, Clark SJ (1996) Cytosine methylation: Quantitation by automated genomic sequencing and GENESCAN(TM) analysis. *BioTechniques*, **21**, 126-133.

Richards EJ (2008) Population epigenetics. *Current Opinion in Genetics & Development*, **18**, 221-226.

Ritchie MG (2007) Sexual selection and speciation. *Annual Review of Ecology, Evolution and Systematics*, **38**, 79-102.

Ritchie MG, Hamill RM, Graves JA, *et al*. (2007) Sex and differentiation: population genetic divergence and sexual dimorphism in Mexican Goodeid fish. *Journal of Evolutionary Biology*, **20**, 2048-2055.

Ritchie MG, Webb SA, Graves JA, Magurran AE, Macías Garcia C (2005) Patterns of speciation in endemic Mexican Goodeid fish: sexual conflict or early radiation? *Journal of Evolutionary Biology*, **18**, 922-929.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365-386.

Sandelin A, Wasserman WW, Lenhard B (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Research*, **32**, 249-252.

Sang XP, Curran MS, Wood AW (2008) Paracrine insulin-like growth factor signaling influences primordial germ cell migration: In vivo evidence from the zebrafish model. *Endocrinology*, **149**, 5035-5042.

Singh KK, Shukla PC, Schmidtke J (2008) Conservation of 5'-upstream region of the *FBN1* gene in primates. *European Journal of Human Genetics*, **16**, 869-872.

Smits G, Mungall AJ, Griffiths-Jones S, *et al*. (2008) Conservation of the *H19* noncoding RNA and *H19-IGF2* imprinting mechanism in therians. *Nature Genetics*, **40**, 971-976.

Stadler PF (2010) Evolution of the long non-coding RNAs *MALAT1* and *MEN* beta/epsilon. *Advances in Bioinformatics and Computational Biology*, **6268**, 1-12.

Tate PH, Bird AP (1993) Effects of DNA methylation on DNA-binding proteins and gene expression. *Current Opinion in Genetics & Development*, **3**, 226-231.

Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y (2003) Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Research*, **13**, 382-390.

Tykocinski ML, Max EE (1984) CG dinucleotide clusters in MHC genes and in 5' demethylated genes. *Nucleic Acids Res*, **12**, 4385-4396.

Van Doorn GS, Edelaar P, Weissing FJ (2009) On the origin of species by natural and sexual selection. *Science*, **326**, 1704-1707.

Vrana PB (2007) Genomic imprinting as a mechanism of reproductive isolation in mammals. *Journal of Mammalogy*, **88**, 5-23.

Wang ZQ, Fung MR, Barlow DP, Wagner EF (1994) Regulation of embryonic growth and lysosomal targeting by the imprinted *Igf2*/*Mpr* Gene. *Nature*, **372**, 464-467.

Webb SA, Graves JA, Macías Garcia C, *et al*. (2004) Molecular phylogeny of the livebearing Goodeidae (Cyprinodontiformes). *Molecular Phylogenetics and Evolution*, **30**, 527-544.

Weber M, Hellmann I, Stadler MB, *et al*. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, **39**, 457-466.

Wood AW, Duan CM, Bern HA (2005) Insulin-like growth factor signaling in fish. *International Review of Cytology*, **243**, 215-285.

Yan MD, Hong CC, Lai GM, *et al*. (2005) Identification and characterization of a novel gene *Saf* transcribed from the opposite strand of *Fas*. *Human Molecular Genetics*, **14**, 1465-1474.

Yeo GW, Van Nostrand EL, Liang TY (2007) Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genetics*, **3**, 814-829.

Zhu LC, Zhang Y, Zhang W, *et al*. (2009) Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*, **10**, 1-12.

Zou SM, Kamei H, Modi Z, Duan CM (2009) Zebrafish *IGF* genes: Gene duplication, conservation and divergence, and novel roles in midline and notochord development. *PLoS ONE*, **4**, 1-12.

CHAPTER 6:

GENERAL DISCUSSION

In this thesis I examined the genetic, transcriptomic and epigenetic variation between divergent populations, in the context of ecological speciation. Ecological speciation occurs through the adaptation of populations to different environments and the consequent evolution of reproductive isolation. Recent shifts in speciation research emphasise the role of ecological speciation and include new ideas on the genetic basis of speciation, such as divergence hitchhiking (Feder & Nosil 2010). Evidence now suggests that the targets of selection may encompass genetic variation in non-coding DNA (Jones *et al*. 2012) and that phenotypic plasticity could play an important role in adaptation and speciation (Thibert-Plante & Hendry 2011). This shift in thinking does not replace the traditional geographical based models of speciation; rather they represent a change in emphasis on the evolutionary processes that create species.

D. mojavensis is a useful species with which to examine ecological speciation, due to its strong ecological links to its host plant. Two populations of *D. mojavensis* display premating reproductive isolation between them that is influenced by the host cactus they utilise, and these species are thought to be incipient species (Brazner & Etges 1993; Etges 1992; Markow 1991; Pfeiler *et al*. 2009). In Chapter 2 I examined two genes that had been previously identified as candidate genes underlying this sexual isolation, finding no evidence for functional genetic divergence between the Baja and Mainland populations. In Chapter 3 I characterised the scenario of *D. mojavensis* population differentiation, inferring both the order of population divergence and timing of these splits in order to place into historical context the ecological and functional genetic data that exists for this species. Chapter 4 examined gene expression differences associated with host plant plasticity, which might be an important aspect of ecological speciation, and linked this plasticity to mating success. The final chapter of my thesis examined divergent populations of *G. multiradiatus* for

evidence of genomic imprinting in the *Igf2* gene associated with sexual conflict, finding little evidence of population-specific patterns, but discovering the presence of DNA methylation, as well as highly conserved non-coding regions that might be linked to gene expression regulation.

The phylogeographic context of incipient ecological speciation

Inferring the demographic context of a speciation event is important to understanding the evolutionary processes underlying that event. *D. mojavensis* is well-studied in its ecology and the system is now recognised for its potential in investigating ecological speciation. As such the functional genetics and genomics of the Baja and Mainland divergence are receiving increasing attention (e.g. Matzkin 2012; Matzkin & Markow 2009; Matzkin *et al*. 2006). However, to understand the functional genetic divergence and to put into context the ecological data gathered over the years, it is important to have some idea of the order of population divergence, how long ago these populations diverged and the levels of gene flow between them. Knowing the geographic region in which *D. mojavensis* originated tells us which population we might expect to discover standing genetic variation that could be important to the evolution of derived populations. Further, knowledge on the timing of population divergence provides an idea of the length of time it took to reach the level of reproductive isolation observed. Lastly, levels of gene flow can help to infer whether the populations diverged under heavy levels of migration, or whether geographic separation allowed them to evolve in isolation. Thus phylogeographic analyses provide the historical and geographical context for the interpretation of the evolutionary and ecological divergence of incipient species.

In Chapter 3 of this thesis I examined the phylogeography of the Baja, Mainland and Mojave populations, inferring the order of population divergence and finding little gene flow accompanying the divergence of all three populations. Using multiple, independent X-linked loci from introns in a model-based framework, I inferred several demographic parameters using an isolation-with-migration model. The strength of this analysis was the use of multiple independent loci, which represent separate sources of information about the species tree (Garrick *et al*. 2010; Hey 2005), and a framework that allowed the examination of many possible different models. This framework consisted of pairwise comparisons of the Baja, Mainland and Mojave populations in order to infer the order of population divergence, allowing a subset of three-population models to be examined for the most likely scenario of divergence. The ranking of models showed that those with very little gene flow between all populations were most likely, although a low level of gene flow between the ancestral Mainland/Mojave population and the Baja population could not be ruled out. The most likely model demonstrated that the Baja and ancestral Mainland/Mojave population diverged around 230-270 Kya, with this ancestral group later splitting into the contemporary Mainland and Mojave populations 117-135 Kya.

These two divergence estimates coincided with two interglacial periods, demonstrating that the Baja-Mainland divergence most likely occurred in allopatry, separated by the Gulf of Mexico. Pleistocene climatic oscillations occurred from around 900 Kya onwards, with a periodicity of approximately 100,000 years (Kraaijeveld & Nieboer 2000). The Holstein interglacial period occurred from 230-215 Kya, with sea levels rising dramatically during that period. The first population divergence within *D. mojavensis* occurred around this time, with rising sea levels isolating mainland Mexico following an invasion of the mainland from the Baja

peninsula. The Eemian interglacial (123-115 Kya) followed, coinciding with the divergence of the Mojave and Mainland populations. This last divergence is most likely due to range reductions and fragmentation of the cactus host plant in the Mojave Desert (barrel cactus, *F. cylindraceous*), due to increased temperatures drying out the region (Axelrod 1983).

Evidence suggests that *D. mojavensis* speciated from *D. arizonae* approximately 2.4 million years ago, with the *D. mojavensis* populations diverging more recently (Matzkin & Eanes 2003). Mexican *Drosophila* species are thought to have moved westwards across Mexico over evolutionary time, speciating as they went. The *D. mojavensis*/*D. arizonae* ancestor most likely diverged due to the splitting of the Baja peninsula from mainland Mexico 3-5 million years ago, giving rise to *D. mojavensis* on the Baja California peninsula (Nason *et al*. 2002; Ruiz *et al*. 1990). Later, *D. mojavensis* colonised mainland Mexico (Matzkin 2004) during a period of low sea level, with the Pleistocene glacial cycles causing population divergence and eventual geographical isolation. Thus geological and climatic extrinsic barriers have contributed to the contemporary distribution of *D. mojavensis*. Importantly, much evidence suggests that the Baja geographic region is where *D. mojavensis* originated, and that the colonisation of mainland Mexico and a host shift was followed by eventual isolation around 230-270 Kya. Since then derived Mainland and ancestral Baja populations evolved a level of premating isolation (Etges *et al*. 2007; Etges *et al*. 2010; Etges *et al*. 2009; Markow 1991), mediated by divergent CHCs and courtship songs and accompanied by a suite of changes in life history traits (Etges 1990; Etges & Heed 1987; Etges 1989, 1993).

Speciation genes

Important to understanding how ecological speciation proceeds is the characterisation of its molecular basis. This includes the genetic basis of speciation (traditional speciation genes) as well as gene expression changes (at regulatory loci) and possibly even epigenetic mechanisms that might underlie phenotypes. The identification of candidate speciation genes can be challenging and this is particularly true where a species displays extensive phenotypic plasticity. For example, QTL analyses are sensitive to plasticity because they associate phenotypes with genetic markers, and unexpected plasticity (i.e. gene expression differences causing trait variation) might decrease this association. Phenotypic plasticity was at one time viewed as noise that was of little importance to evolution, however it is now realised that phenotypic plasticity might be an important evolutionary process (Fitzpatrick 2012).

In Chapter 2 I examined two candidate genes for population-specific functional variation that might explain CHC differences and reproductive isolation between the Baja and Mainland populations of *D. mojavensis*. *Desat1* and *desat2* are Δ9 desaturase genes that have previously been implicated in the reproductive isolation of *D. melanogaster* populations (Dallerac *et al*. 2000; Takahashi *et al*. 2001). However, little genetic variation was discovered between the *D. mojavensis* populations across both genes, and no mutations were observed that changed the amino acid sequence of the proteins. Thus no functional changes were seen between populations in *desat1* or *desat2*, indicating that these genes might not play a role in the divergent reproductive phenotypes that underlie population sexual isolation. This conclusion was also supported by results from Chapter 4, in which no Δ9 desaturases were differentially or alternatively expressed between cactus hosts or involved in the mating success of males. However, Chapter 4 did not include examination of Baja

males, and Chapter 2 did not cover the promoter region of *desat1*, and thus it remains possible that genetic variation in regulatory features between the Baja and Mainland populations in the *desat1* promoter, might lead to population-specific expression of this gene. *Desat1* expression can be complex, with recent work showing it can produce several different transcripts involved in coordinating the production and perception of courtship pheromones, and that it has rapidly evolving regulatory regions (Bousquet & Ferveur 2012; Bousquet *et al*. 2012). Thus it is possible that the Mainland and Baja populations of *D. mojavensis* might be divergent in *desat1* regulatory features. However, the evidence in Chapter 4 suggests that gene expression related to pheromone production is not involved in cactus plasticity or male mating success, although several genes associated with olfaction and gustation were seen to be plastic across cacti. The role of the desaturases in adaptation and behavioural isolation between *D. melanogaster* populations has also been questioned, with evidence for (Fang *et al*. 2002; Greenberg *et al*. 2003) and against (Coyne & Elwyn 2006; Grillet *et al*. 2012) their involvement. These opposing results highlight the difficulty in identifying genes that underlie processes such as adaptation and speciation. Even when there is evidence for a particular adaptive phenotype, and candidate genes, it can be difficult to narrow down the causative variation linked to adaptation and reproductive isolation. This is where multiple lines of independent evidence can be an advantage, especially when performed genome-wide to obtain all sets of selected loci. Further, examining both genetic and expression variation together is advantageous, especially for recently diverged populations or species because selection may be able to act more efficiently on *cis*-regulatory regions than coding, as pleiotropy can constrain the evolution of protein coding regions (Jones *et al*. 2012; Wray 2007).

Gene expression plasticity

The role of phenotypic plasticity in evolution has previously been contentious with some debate as to whether plasticity contributes to or inhibits adaption and speciation (Fitzpatrick 2012). However, recent theoretical work (Thibert-Plante & Hendry 2011) and empirical studies (Levine *et al*. 2011) demonstrate that phenotypic plasticity may promote colonization and adaptation to new environments. In Chapter 4 I examined gene expression patterns according to early development (host cactus rearing) in adult Mainland male flies, and male mating success. Associating host cactus and mating success gene expression aimed to identify the functions of genes expressed according to cactus-dependent mating success. The cactus influence on male mating success in *D. mojavensis* has been demonstrated extensively (Etges *et al*. 2007; Etges *et al*. 2010; Etges *et al*. 2009) and the influence of the rearing substrate is thought to occur early in development, where flies develop in the necrotic tissue.

The results of Chapter 4 showed that cactus-dependent mating success was associated with the expression of genes functioning in translation, transcription and neurological development. The nervous system is thought to be an important aspect of courtship and mating in *Drosophila*. Numerous studies have linked genes involved in neurological development and function to courtship behaviour and success in *Drosophila* males (Ditch *et al*. 2005; Finley *et al*. 1998; Finley *et al*. 1997; Gleason 2005; Grosjean *et al*. 2008; Sokolowski 2001; Villella *et al*. 1997; Villella & Hall 2008). Further, neurological genes are thought to be involved with courtship song and pheromone production in males (Ferveur *et al*. 1995; Manoli *et al*. 2005; Peixoto & Hall 1998) and their expression is particularly associated with neural tissues, which in insects have been implicated in control of associative learning and various aspects of behaviour (O'Dell *et al*. 1995; Zanini *et al*. 2012). In Chapter 4 several genes

associated with olfactory and chemosensory behaviour and cognition were differentially expressed across cacti (*Obp99A*, *Or83A*, *Gr94A* and *drk*). Thus it seems that cactus specific gene expression includes the sensing and parsing of cues from the host cactus environment. This might be related to host plant choice and preference, which often involves olfactory cues in insects and might play a role in environment-specific reproductive success (Cunningham 2012; Grosjean *et al*. 2011). Further, gustatory (and potentially olfactory) genes have neural links to the reproductive organs, providing a mechanistic link between sensing the external environment and resulting effects on reproductive success (Park & Kwon 2011). However, evidence shows that the Baja population, and majority of Mainland subpopulations, prefer agria cactus to organ pipe as their hosts, suggesting that a host shift to organ pipe has not led to substantial changes in host preference (Newby & Etges 1998). Our understanding of how behaviour evolves is limited especially in terms of host plant use and chemosensory systems, integral to this is determining the role of the central and peripheral nervous systems in ecological adaptation and the evolution of behavioural reproduction isolation (Cande *et al*. 2012).

The presence of organ pipe cactus on the Baja peninsula and the extensive genotype-by-environment interactions between the Baja and Mainland populations suggests that phenotypic plasticity could have played a role in their divergence. *D. mojavensis* originated on the Baja California peninsula after tectonic activity opened up the Gulf of Mexico 3-5 million years ago (Ruiz 1990; Matzkin 2004; Nason *et al*. 2002). Organ pipe is also present in Baja California, located in the south of the peninsula. Thus it is likely that *D. mojavensis* has encountered organ pipe cactus throughout its evolution, but adapted to the most prevalent environment. However, the presence of even low levels of gene flow between cactus environments is likely to

have resulted in selection for some degree of a plastic response, if given enough time to evolve (Yampolsky *et al*. 2012) and as long as changes in the environment were somewhat predictable, as is the case with migration between environments (Scheiner 1998). The invasion of mainland Mexico during a period of low sea level would have led to increased gene flow between ecological environments, maintaining levels of plasticity. Rising sea levels eventually isolated these two populations around 230,000 years ago, leading to predominant use of organ pipe cactus on the mainland and aided by adaptive plastic responses evolved in the ancestral Baja population. Plasticity was maintained in the Mainland population for a number of life history and reproductive traits, with selection acting on trait means across environments (genetic accommodation). Selection in the Mainland population increased thorax size, development times and egg to adult viability (Etges 1993; Etges *et al*. 2010) and led to the evolution of shorter long-interpulse intervals (L-IPI) in male courtship song (Etges et al, 2006), and a female preference for L-IPI (Etges *et al*. 2007). Further, increased amounts of three alkadiene components of CHCs evolved in the Mainland population (Etges & Jackson 2001; Stennett & Etges 1997), and a low but significant level of reproductive isolation (Etges *et al*. 2007; Etges *et al*. 2010; Etges *et al*. 2009; Markow 1991). The maintenance of plasticity was likely due to the existence of a number of different host cactus species in the region, as well as a small patch of agria at Punta Onah, presumably with few associated costs and limits to plasticity. Thus it is possible that phenotypic plasticity played a key role in the adaptation of *D. mojavensis* populations to organ pipe cactus and their incipient speciation.

The role of epigenetics

The role of epigenetics in evolution has been a controversial topic in the last few years, with proponents (Richards *et al*. 2010) and opponents (Dickins & Rahman 2012) debating its significance to evolutionary studies. In this thesis I have examined or encountered epigenetics in three chapters; Chapter 2, Chapter 5 and Chapter 4 and thus it seems pertinent to consider what role epigenetic marks might play in evolution. Epigenetic marks control gene expression through the regulation of heterochromatin or through more direct regulation, such as interference with transcription and translation machinery (Bossdorf *et al*. 2008; Jaenisch & Bird 2003; Ohlsson *et al*. 2001). Epigenetic mechanisms play a key developmental role by regulating gene expression in somatic cells, however germ line cells can also be epigenetically marked, for example through genomic imprinting. Although epigenetic variation can be environmentally induced, stably inherited and underlie quantitative trait variation (Becker *et al*. 2011; Johannes *et al*. 2009; Richards *et al*. 2010), the role of epigenetics in evolution is still under debate. One major factor contributing to the controversy of this role is that epigenetics is often viewed as being a Larmarckian process (Jablonka & Lamb 1995) and is thus dismissed because it opposes the neo-Darwinian view of natural selection, and genetic evolution, which is central to the Modern Synthesis (Haig 2007). First, an important aspect of neo-Darwinism and the Modern Synthesis is that evolution is not 'directed' and thus because epigenetic variation can be induced and inherited at non-random genomic locations, it is seen as Lamarckian. Second, epigenetic variation is not seen as a primary driver of evolution as, say, novel mutations can be. Last, epigenetics is defined as independent of genetics and thus, under the gene centric view of the Modern Synthesis, cannot be involved in genetic evolution. However, these criticisms come from a lack of

knowledge about epigenetic mechanism and function and, far from being Lamarckian, epigenetic marks may evolve under Darwinian evolution, and have an important role in phenotypic evolution (Richards, 2006; 2008). The above three aspects of epigenetics are explored below in turn.

The environment plays a dual role in evolution, both in creating a phenotype alongside genotype (through developmental plasticity), and as acting as a selection pressure on that phenotype (leading to adaptation). Phenotypic plasticity can evolve because it can be adaptive and it is likely that many plastic responses have an epigenetic basis (Jaenisch & Bird 2003; Johnson & Tricker 2010; Levine *et al*. 2011; Richards *et al*. 2010). Similarly, epigenetic mechanisms are likely to have evolved because they are adaptive. The environment can induce epigenetic marks anew every generation, and these often occur in specific regions of the genome (e.g. differentially methylated regions, DMRs). Further, if such marks can be induced in the germ line, if a mechanism exists by which external cues are parsed to the gametes, then this induced heritable epigenetic mark would 'program' future offspring for their environment. Care must be taken in testing this last idea because exposure of each new generation to an environment (e.g. maternal effects) will produce the same phenotypic results as induced epigenetic inheritance. However, there is now some evidence in mammals (Daxinger & Whitelaw 2012) and *Drosophila* (Seong *et al*. 2011) that induced epigenetic inheritance occurs. Further, theoretical models suggest that such a mechanism could provide an adaptive advantage in stochastic environments, providing a link between short-term individual adaptive responses and long-term evolutionary change (Jablonka *et al*. 1995). However, more empirical work is necessary to examine this idea further. It should be noted that such a mechanism is not 'directed' evolution as such, but more akin to phenotypic plasticity in that it is an

evolved mechanism to deal with environmental change. Thus environmental induction of epigenetic variation, whether in the germline or soma, still arose through Darwinian processes.

Evidence now suggests that epimutations (newly derived epigenetic marks) can arise stochastically, be heritable, underlie quantitative traits, have similar mutation rates to genetic mutations and contribute to phenotypic evolution. The majority of empirical studies that examine epigenetics focus on DNA methylation, which has been extensively examined in plants and where much progress has been made in understanding the epigenetic basis of traits. Evidence suggests that novel epimutations can occur spontaneously across generations, in somatic cells with aging, in the male germline during meiosis (de Boer *et al*. 2010) and through mistakes during reprogramming, post-fertilization (Schmitz *et al*. 2011). Schmitz *et al* (2011) demonstrated spontaneously occurring DNA methylation in *Arabidopsis thaliana*, at more than 100,000 CpG positions occurring over 30 generations, the majority of which were meiotically stable. Johannes *et al*, (2009) discovered epialleles that were stable over at least 8 generations and contributed to heritable variation in complex traits. Similar to DNA allelic variation, epialleles are defined as 'alternative chromatin states at a given locus defined with respect to individuals in a population at a given time point and tissue type' (Johannes *et al*. 2008). Epimutations at individual positions in *A. thaliana* are much more frequent than DNA mutations, with the lower bound of epimutation rates being 4.46×10^{-4} methylation polymorphisms per CG site per generation (Schmitz *et al*. 2011), compared with genetic mutation rates of 7×10^{-9} base substitutions per site per generation in the same *A. thaliana* lines (Ossowski *et al*. 2010). However, epimutation rates in contiguous regions of methylation, which are likely to be functional (e.g. DMRs), are more in the range of DNA mutational change

(Becker *et al*. 2011). Becker *et al* (2011) found 249 DMRs that tended to be more stable than individual positions over the 30 generations examined. This indicates that epimutations occur as frequently as genetic mutations and are stable at certain key functional genomic positions. Lastly, this has been shown to be true in natural populations of *A. thaliana*, with methylation variation causing heritable phenotypic variation (Roux *et al*. 2011), and Vaughn *et al* (2007) found that natural populations of *A. thaliana* demonstrated high levels of epimutation polymorphism among ecotypes. Thus there is much evidence for the importance of epigenetic novelty in Darwinian evolution, i.e. heritable epigenetic variation that might be under selection. Further, such evidence is not limited to plants, with increasing numbers of studies demonstrating epigenetic variation in mammals (Daxinger & Whitelaw 2012), insects (Seong *et al*. 2011) and nematodes (Greer *et al*. 2011).

Ultimately, epigenetic variation is linked to genetic variation to some degree because DNA underlies the proteins that create marks, and epigenetic marks occur with, and depend on DNA, RNA or proteins. Thus critics of the role of epigenetics in evolution might say that it does not have a role in, or influence on, genetic evolution. However, Richards (2006) set out three categories of epigenetic mark, based on their dependence on genotype. These are; (i) epigenetic variation entirely dependent on genetic variation, (ii) facilitated epigenetic variation in which the genotype directs epigenetic marks in a probabilistic but not deterministic manner, and (iii) epigenetic variation generated by stochastic events, largely independent from the DNA sequence. It is clear that all epigenetic marks rely on genetic mechanisms to some degree, yet the genotype does not always predict the epigenotype (Richards 2008). However, recent theoretical models have examined epigenetics in terms of the genes that control the expression of phenotypic variance but do not affect the mean of traits (Carja &

Feldman 2012; Feinberg & Irizarry 2010). Feinberg and Irizarry (2010) found evidence for genetic variants that change the variance but not the mean of phenotypes, and that this is mediated epigenetically. They empirically examined DMRs and variably methylated regions (VMRs) in mice and found that VMRs were often located near DMRs, indicating that VMRs might evolve into DMRs over time. Further, they discovered that DMRs across species differed in their underlying DNA sequence (loss or gain of CpG dinucleotides), indicating that functional methylation relies heavily on genetic variation for CpG motifs. Feinberg and Irizarry (2010) then simulated the effect of selection on genetic loci that can modify the variance of phenotypes (through epigenetic variation), without influencing the mean of a trait, across fluctuating environments and showed that such a mechanism provides an adaptive advantage. Carja and Feldman (2012) recently extended this model over a longer evolutionary timescale, showing that initially, a fluctuating environment does increase phenotypic variation, however this increase reaches equilibrium at around 10,000 generations. These studies demonstrate the importance of interacting genetic and epigenetic variation over relatively short evolutionary time periods.

Epigenetic marks may feedback into genetic evolution through changes in mutation rates or by shielding DNA from selection, influencing mutational change and contributing to long-term evolution. Epigenetic marks might influence evolution in several ways, such as the suppression of transposable elements or an effect on recombination (Richards 2008). Additionally, 5-methylcytosine is subject to spontaneous deamination, turning a methylated C into a T, at ~12 times higher a rate than traditional DNA mutations (Laird & Jaenisch 1994; Sved & Bird 1990) and CpG depletion can occur over generations, in regions that are targeted for methylation (Flores & Amdam 2011). Ossowski *et al* (2010) examined mutation accumulation

lines in *A. thaliana* and found that there were more G:C > A:T transition mutations than expected by chance. They demonstrated that methylated sites had a higher probability of mutating in this manner, concluding that a combination of UV-induced mutagenesis and deamination of methylated cytosines accounted for these mutations. The production of facultative heterochromatin involves several histone cores that wrap DNA, preventing gene expression by blocking the attachment of transcription machinery (Grewal & Moazed 2003). This process shields silenced regions from the action of selection, allowing mutations to accumulate, in much the same way as heterochromatin accumulates transposable elements (Dimitri & Junakovic 1999). Once heterochromatin becomes euchromatin again, accumulated mutations in coding regions might lead to a loss of function, on which selection would act. Thus epigenetic variation can be linked to genetic variation to different degrees and even when linked, epigenetic variation might feed back into the genome.

Considering the role of epigenetics in evolution does not challenge the Darwinian view of evolution. Epigenetic mechanisms can play a role in phenotypic plasticity, in terms of developmental environmental responses within and possibly across generations. Further, epigenetic novelty can arise in much the same way mutations do, be heritable and contribute to quantitative characters and epigenetic marks can feedback into the genome, influencing genetic evolution. In Chapter 4, gene expression associated with epigenetic mechanisms was demonstrated, with a role for both methylation (predominantly of proteins and RNA) and histone modifications across cactus hosts and male mating success in *D. mojavensis*. A recent study found that chromatin remodelling factors (such as histone modifying genes) were involved in plasticity and genotype-by-environment interactions between tropical and temperate populations of *D. melanogaster* (Levine *et al.* 2011). Epigenetic

mechanisms in both these cases seem to underlie a plastic response to the environment and to involve non-DNA methylation mechanisms. *Drosophila* are part of the '*Dnmt2* only' group of species, that lack the core DNA methyltransferase genes (Krauss & Reuter 2011). Thus in these species epigenetic mechanisms other than DNA methylation seem to be utilised for important functions, at least in terms of phenotypic plasticity.

The future investigation of the role of epigenetics in evolution should try to partition the roles of different epigenetic mechanisms and assess how they might contribute to evolutionary change. For instance, what is the role of mechanisms that have evolved to allow adaptive environmentally induced variation, compared to selection on novel epimutations? What is the role of genetic dependent compared to independent epigenetic variation? How much influence does feedback from epigenetic marks actually have on genetic variation and evolution? There are many exciting new studies that are beginning to examine epigenetics in more detail. For example, a recent study implicates DNA methylation in reversible epigenetic changes underlying honeybee castes, suggesting a role for epigenetics in behaviour (Herb *et al*. 2012). There is much we do not know about epigenetic mechanisms and their role in evolution and more empirical studies are required to examine this role, across a range of different organisms.

The molecular basis of ecological speciation

Ecological speciation represents a reframing of the role of ecology in evolution and includes new theories on how speciation occurs at the genetic level. However, important to understanding how speciation proceeds is to identify the targets of natural selection. From our increasing knowledge on the structure of genomes and

how they operate, it is clear that the molecular basis of phenotypes is likely to be complex. For example, the human ENCODE project recently attempted to summarise the structure of the human genome, in which only 2.94% codes for protein coding genes. In total, 62% of the genome was found to be non-coding RNA, both long and short RNAs that can regulate gene expression, with the majority of these located inside of introns or near genes (ENCODE project consortium 2012). Further, they identified the genomic positions of histone modifications and DNA methylation and associated these with functional genetic regions, in order to dissect how different epigenetic marks regulate gene expression. Although the ENCODE project is far from providing a final blue print of the human genome (and in fact makes bold over-statements about the data, mostly derived from a liberal definition of 'function'), these results do at least highlight the complexity of genomes meaning that the genomic targets of selection, for even a single phenotype, may be diverse. In particular, regulatory regions might be prominent targets, especially during the initial stages of ecological speciation, because *cis* regulatory mutations allow for a change in phenotype without changing the protein, circumventing functional trade-offs due to pleiotropy and allowing selection to operate more efficiently (Wray 2007). For example Jones *et al* (2012) discovered that 41% of divergent genetic regions between marine and freshwater stickleback ecotypes (*Gasterosteus aculeatus*) were regulatory, with a further 42% likely to be regulatory, and only 17% of changes occurring in coding regions.

Therefore, the key to understanding the molecular basis of ecological speciation is an increased understanding of how genomes are expressed. The improvement of molecular technologies that enable the combination of several lines of evidence (e.g. neutral genetic divergence, functional selective divergence,

expression variation and epigenetic variation) will aid in this goal. In this thesis I have set out the historical context of the divergence of *D. mojavesis* populations, in order to understand the evolution of reproductive isolation between the Baja and Mainland populations. Further, I identified a set of functionally important genes involved in phenotypic plasticity across host cacti and male mating success, and established that genes directly involved in pheromone expression are unlikely to underlie this incipient speciation. Finally, I found a likely role for epigenetic mechanisms in phenotypic plasticity, which in *Drosophila* is unlikely to involve DNA methylation.

REFERENCES

Axelrod DI (1983) *Paleobotanical history of the western deserts* University of New Mexico Press, Albuquerque.

Becker C, Hagmann J, Muller J, *et al*. (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature*, **480**, 245-U127.

Bossdorf O, Richards CL, Pigliucci M (2008) Epigenetics for ecologists. *Ecology Letters*, **11**, 106-115.

Bousquet F, Ferveur JF (2012) *desat1*: A Swiss army knife for pheromonal communication and reproduction? *Fly (Austin)*, **6**, 102-107.

Bousquet F, Nojima T, Houot B, *et al*. (2012) Expression of a desaturase gene, *desat1*, in neural and nonneural tissues separately affects perception and emission of sex pheromones in *Drosophila*. *Proceedings of the National Academy of Sciences, USA*, **109**, 249-254.

Brazner JC, Etges WJ (1993) Pre-mating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. II. Effects of larval substrates on time to copulation, mate choice and mating propensity. *Evolutionary Ecology*, **7**, 605-624.

Cande J, Prud'homme B, Gompel N (2012) Smells like evolution: the role of chemoreceptor evolution in behavioral change. *Current Opinions in Neurobiol*ogy, **23**, 1-7.

Carja O, Feldman MW (2012) An equilibrium for phenotypic variance in fluctuating environments owing to epigenetics. *Journal of the Royal Society Interface*, **9**, 613-623.

Coyne JA, Elwyn S (2006) Does the *desaturase-2* locus in *Drosophila melanogaster* cause adaptation and sexual isolation? *Evolution*, **60**, 279-291.

Cunningham J (2012) Can mechanism help explain insect host choice? *Journal of Evolutionary Biology*. **doi: 10.1111/j.1420-9101.2011.02435.x**

Dallerac R, Labeur C, Jallon JM, *et al*. (2000) A Delta 9 desaturase gene with a different substrate specificity is responsible for the cuticular diene hydrocarbon polymorphism in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences, USA*, **97**, 9449-9454.

Daxinger L, Whitelaw E (2012) Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nature Reviews Genetics*, **13**, 153-162.

de Boer P, Ramos L, de Vries M, Gochhait S (2010) Memoirs of an insult: sperm as a possible source of transgenerational epimutations and genetic instability. *Molecular Human Reproduction*, **16**, 48-56.

Dickins TE, Rahman Q (2012) The extended evolutionary synthesis and the role of soft inheritance in evolution. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **279**, 2913-2921.

Dimitri P, Junakovic N (1999) Revising the selfish DNA hypothesis: New evidence on accumulation of transposable elements in heterochromatin. *Trends in Genetics*, **15**, 123-124.

Ditch LM, Shirangi T, Pitman JL, *et al*. (2005) *Drosophila* retained/dead ringer is necessary for neuronal pathfinding, female receptivity and repression of fruitless independent male courtship behaviors. *Development*, **132**, 155-164.

ENCODE project consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-73.

Etges W (1990) Direction of life history evolution in *Drosophila mojavensis*. In: *Ecological and Evolutionary Genetics of Drosophila*, pp. 37-56.

Etges W, Heed W (1987) Sensitivity to larval density in populations of *Drosophila mojavensis*: Influences of host plant variation on components of fitness. *Oecologia*, **71**, 375-381.

Etges WJ (1989) Evolution of developmental homeostasis in *Drosophila mojavensis*. *Evolutionary Ecology*, **3**, 189-201.

Etges WJ (1992) Premating isolation is determined by larval substrates in cactophilic *Drosophila mojavensis*. *Evolution*, **46**, 1945-1950.

Etges WJ (1993) Genetics of host-cactus response and life-history evolution among ancestral and derived populations of cactophilic *Drosophila mojavensis*. *Evolution*, **47**, 750-767.

Etges WJ, de Oliveira CC, Gragg E, *et al*. (2007) Genetics of incipient speciation in *Drosophila mojavensis*. I. Male courtship song, mating success, and genotype x environment interactions. *Evolution*, **61**, 1106-1119.

Etges WJ, De Oliveira CC, Noor MAF, Ritchie MG (2010) Genetics of incipient speciation in *Drosophila mojavensis*. III. Life history divergence in allopatry and reproductive isolation. *Evolution*, **64**, 3549-3569.

Etges WJ, de Oliveira CC, Ritchie MG, Noor MAF (2009) Genetics of incipient speciation in *Drosophila mojavensis*. II. Host plants and mating status influence cuticular hydrocarbon QTL Expression and G × E Interactions. *Evolution*, **63**, 1712-1730.

Etges WJ, Jackson LL (2001) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. VI. Epicuticular hydrocarbon variation in *Drosophila mojavensis* cluster species. *Journal of Chemical Ecology*, **27**, 2125-2149.

Fang S, Takahashi A, Wu CI (2002) A mutation in the promoter of *desaturase 2* is correlated with sexual isolation between *Drosophila* behavioral races. *Genetics*, **162**, 781-784.

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729-1747.

Feinberg AP, Irizarry RA (2010) Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences, USA*, **107**, 1757-1764.

Ferveur JF, Stortkuhl KF, Stocker RF, Greenspan RJ (1995) Genetic feminization of brain structures and changed sexual orientation in male *Drosophila*. *Science*, **267**, 902-905.

Finley KD, Edeen PT, Foss M, *et al*. (1998) Dissatisfaction encodes a tailless-like nuclear receptor expressed in a subset of CNS neurons controlling *Drosophila* sexual behavior. *Neuron*, **21**, 1363-1374.

Finley KD, Taylor BJ, Milstein M, McKeown M (1997) dissatisfaction, a gene involved in sex-specific behavior and neural development of Drosophila melanogaster. *Proceedings of the National Academy of Sciences, USA*, **94**, 913-918.

Fitzpatrick BM (2012) Underappreciated consequences of phenotypic plasticity for ecological speciation. *International Journal of Ecology*, **doi:10.1155/2012/256017**.

Flores KB, Amdam GV (2011) Deciphering a methylome: what can we read into patterns of DNA methylation? *The Journal of Experimental Biology*, **214**, 3155-3163.

Garrick RC, Caccone A, Sunnucks P (2010) Inference of population history by coupling exploratory and model-driven phylogeographic analyses. *International Journal of Molecular Sciences*, **11**, 1190-1227.

Gleason JM (2005) Mutations and natural genetic variation in the courthship song of *Drosophila*. *Behavior Genetics*, **35**, 265-277.

Greenberg AJ, Moran JR, Coyne JA, Wu CI (2003) Ecological adaptation during incipient speciation revealed by precise gene replacement. *Science*, **302**, 1754-1757.

Greer EL, Maures TJ, Ucar D, *et al*. (2011) Transgenerational epigenetic inheritance of longevity in *Caenorhabditis elegans*. *Nature*, **479**, 365-371.

Grewal SIS, Moazed D (2003) Heterochromatin and epigenetic control of gene expression. *Science*, **301**, 798-802.

Grillet M, Everaerts C, Houot B, *et al*. (2012) Incipient speciation in *Drosophila melanogaster* involves chemical signals. *Scientific Reports*, **2**, 224.

Grosjean Y, Grillet M, Augustin H, Ferveur JF, Featherstone DE (2008) A glial amino-acid transporter controls synapse strength and courtship in *Drosophila*. *Nature Neuroscience*, **11**, 54-61.

Grosjean Y, Rytz R, Farine JP, *et al*. (2011) An olfactory receptor for food-derived odours promotes male courtship in *Drosophila*. *Nature*, **478**, 236-240.

Haig D (2007) Weismann rules! OK? Epigenetics and the Lamarckian temptation. *Biology and Philosophy*, **22**, 415-428.

Herb BR, Wolschin F, Hansen KD, *et al*. (2012) Reversible switching between epigenetic states in honeybee behavioral subcastes. *Nature Neuroscience*, **doi:10.1038/nn.3218**.

Hey J (2005) On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology*, **3**, e193.

Jablonka E, Lamb MJ (1995) *Epigenetic Inheritance and Evolution: The Lamarckian Dimension* Oxford University Press, Oxford.

Jablonka E, Oborny B, Molnar I, *et al*. (1995) The adaptive advantage of phenotypic memory in changing environments. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **350**, 133-141.

Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics*, **33**, 245-254.

Johannes F, Colot V, Jansen RC (2008) Epigenome dynamics: A quantitative genetics perspective. *Nature Reviews Genetics*, **9**, 883-890.

Johannes F, Porcher E, Teixeira FK, *et al*. (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genetics*, **5**, e1000530.

Johnson LJ, Tricker PJ (2010) Epigenomic plasticity within populations: Its evolutionary significance and potential. *Heredity*, **105**, 113-121.

Jones FC, Grabherr MG, Chan YF, *et al*. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55-61.

Kraaijeveld K, Nieboer EN (2000) Late Quaternary paleogeography and evolution of arctic breeding waders. *Ardea*, **88**, 193-205.

Krauss V, Reuter G (2011) DNA methylation in *Drosophila*: A critical evaluation. *Progress in Molecular Biology and Translational Science*, **101**, 177-191.

Laird PW, Jaenisch R (1994) DNA methylation and cancer. *Human Molecular Genetics*, **3**, 1487-1495.

Levine MT, Eckert ML, Begun DJ (2011) Whole-genome expression plasticity across tropical and temperate *Drosophila melanogaster* populations from Eastern Australia. *Molecular Biology and Evolution*, **28**, 249-256.

Manoli DS, Foss M, Villella A, *et al*. (2005) Male-specific fruitless specifies the neural substrates of *Drosophila* courtship behaviour. *Nature*, **436**, 395-400.

Markow TA (1991) Sexual isolation among populations of *Drosophila mojavensis*. *Evolution*, **45**, 1525-1529.

Matzkin LM (2012) Population transcriptomics of cactus host shifts in *Drosophila mojavensis*. *Molecular Ecology*. **doi: 10.1111/j.1365-294X.2012.05549.x**

Matzkin LM, Eanes WF (2003) Sequence variation of alcohol dehydrogenase (*Adh*) paralogs in cactophilic *Drosophila*. *Genetics*, **163**, 181-194.

Matzkin LM, Markow TA (2009) Transcriptional regulation of metabolism associated with the increased desiccation resistance of the cactophilic *Drosophila mojavensis*. *Genetics*, **182**, 1279.

Matzkin LM, Watts TD, Bitler BG, Machado CA, Markow TA (2006) Functional genomics of cactus host shifts in *Drosophila mojavensis*. *Molecular Ecology*, **15**, 4635-4643.

Nason JD, Hamrick J, Fleming TH (2002) Historical vicariance and postglacial colonization effects on the evolution of genetic structure in *Lophocereus,* a Sonoran Desert columnar cactus. *Evolution*, **56**, 2214-2226.

Newby BD, Etges WJ (1998) Host preference among populations of *Drosophila mojavensis* (Diptera: Drosophilidae) that use different host cacti. *Journal of Insect Behavior*, **11**, 691-712.

O'Dell KM, Armstrong JD, Yang MY, Kaiser K (1995) Functional dissection of the *Drosophila* mushroom bodies by selective feminization of genetically defined subcompartments. *Neuron*, **15**, 55-61.

Ohlsson R, Renkawitz R, Lobanenkov V (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in Genetics*, **17**, 520-527.

Ossowski S, Schneeberger K, Lucas-Lledo JI, *et al*. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92-94.

Park JH, Kwon JY (2011) A systematic analysis of *Drosophila* gustatory receptor gene expression in abdominal neurons which project to the central nervous system. *Molecules and Cells*, **32**, 375-381.

Peixoto AA, Hall JC (1998) Analysis of temperature-sensitive mutants reveals new genes involved in the courtship song of *Drosophila*. *Genetics*, **148**, 827-838.

Pfeiler E, Castrezana S, Reed L, Markow T (2009) Genetic, ecological and morphological differences among populations of the cactophilic *Drosophila mojavensis* from southwestern USA and northwestern Mexico, with descriptions of two new subspecies. *Journal of Natural History*, **43**, 923-938.

Richards CL, Bossdorf O, Pigliucci M (2010) What role does heritable epigenetic variation play in phenotypic evolution? *Bioscience*, **60**, 232-237.

Richards EJ (2006) Inherited epigenetic variation - revisiting soft inheritance. *Nature Reviews Genetics*, **7**, 395-401.

Richards EJ (2008) Population epigenetics. *Current Opinion in Genetics & Development*, **18**, 221-226.

Roux F, Colomé-Tatché M, Edelist C, *et al*. (2011) Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. *Genetics*, **188**, 1015-1017.

Ruiz A, Heed W, Wasserman M (1990) Evolution of the Mojavensis cluster of cactophilic *Drosophila* with descriptions of two new species. *Journal of Heredity*, **81**, 30-42.

Scheiner SM (1998) The genetics of phenotypic plasticity. VII. Evolution in a spatially-structured environment. *Journal of Evolutionary Biology*, **11**, 303-320.

Schmitz RJ, Schultz MD, Lewsey MG, *et al*. (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science*, **334**, 369-373.

Seong KH, Li D, Shimizu H, Nakamura R, Ishii S (2011) Inheritance of stress-induced, ATF-2-dependent epigenetic change. *Cell*, **145**, 1049-1061.

Sokolowski MB (2001) *Drosophila*: genetics meets behaviour. *Nature Reviews Genetics*, **2**, 879-890.

Stennett MD, Etges WJ (1997) Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. III. Epicuticular hydrocarbon variation is determined by use of different host plants in *Drosophila mojavensis* and *Drosophila arizonae*. *Journal of Chemical Ecology*, **23**, 2803-2824.

Sved J, Bird A (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the national academy of sciences, USA*, **87**, 4692.

Takahashi A, Tsaur SC, Coyne JA, Wu C-I (2001) The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila*

melanogaster. *Proceedings of the National Academy of Sciences, USA*, **98**, 3920-3925.

Thibert-Plante X, Hendry AP (2011) The consequences of phenotypic plasticity for ecological speciation. *Journal of Evolutionary Biology*, **24**, 326-342.

Vaughn MW, Tanurdzic M, Lippman Z, *et al*. (2007) Epigenetic natural variation in *Arabidopsis thaliana*. *PLoS Biology*, **5**, 1617-1629.

Villella A, Gailey DA, Berwald B, *et al*. (1997) Extended reproductive roles of the fruitless gene in *Drosophila melanogaster* revealed by behavioral analysis of new *fru* mutants. *Genetics*, **147**, 1107-1130.

Villella A, Hall JC (2008) Neurogenetics of courtship and mating in *Drosophila*. *Advances in Genetics*, **62**, 67-184.

Wray GA (2007) The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews Genetics*, **8**, 206-216.

Yampolsky LY, Glazko GV, Fry JD (2012) Evolution of gene expression and expression plasticity in long-term experimental populations of *Drosophila melanogaster* maintained under constant and variable ethanol stress. *Molecular Ecology*, **21**, 4287-4299.

Zanini D, Jallon JM, Rabinow L, Samson ML (2012) Deletion of the *Drosophila* neuronal gene found in neurons disrupts brain anatomy and male courtship. *Genes, Brain and Behavior*. **doi: 10.1111/j.1601-183X.2012.00817.x**

APPENDIX I:

Macías Garcia C, **Smith G**, Zuarth CG, Graves JA, Ritchie MG (2012) Variation in sexual dimorphism and assortative mating do not predict genetic divergence in the sexually dimorphic Goodeid fish *Girardinichthys multiradiatus*. *Current Zoology*, **58**, 440-452.

Current Zoology 58 (3): 440–452, 2012

Variation in sexual dimorphism and assortative mating do not predict genetic divergence in the sexually dimorphic Goodeid fish *Girardinichthys multiradiatus*

C. MACÍAS GARCIA[1], G. SMITH[2], C. GONZÁLEZ ZUARTH[1], J. A. GRAVES[2], M. G. RITCHIE[2*]

[1] Departamento de Ecologia Evolutiva, Instituto de Ecologia, UNAM, A. P. 70-275, C. P. 04510 México DF, México

[2] School of Biology, University of St Andrews, St Andrews, Fife, KY16 9TH, Scotland, UK

Abstract Sexual dimorphism is often used as a proxy for the intensity of sexual selection in comparative studies of sexual selection and diversification. The Mexican Goodeinae are a group of livebearing freshwater fishes with large variation between species in sexual dimorphism in body shape. Previously we found an association between variation in morphological sexual dimorphism between species and the amount of gene flow within populations in the Goodeinae. Here we have examined if morphological differentiation within a single dimorphic species is related to assortative mating or gene flow between populations. In the Amarillo fish *Girardinichthys multiradiatus* studies have shown that exaggerated male fins are targets of female preferences. We find that populations of the species differ in the level of sexual dimorphism displayed due to faster evolution of differences in male than female morphology. However, this does not predict variation in assortative mating tests in the laboratory; in fact differences in male morphology are negatively correlated with assortative mating. Microsatellite markers reveal significant genetic differences between populations. However, gene flow is not predicted by either morphological differences or assortative mating. Rather, it demonstrates a pattern of isolation by distance with greater differentiation between watersheds. We discuss the caveats of predicting behavioural and genetic divergence from so-called proxies of sexual selection [*Current Zoology* 58 (3): 440–452, 2012].

Keywords Sexual dimorphism, Assortative mating, Genetic distance, Speciation, Viviparous fish

Speciation is the cornerstone of biological diversity and studies of the processes that contribute to the divergence of populations and species are central to understanding speciation. The study of speciation has probably never been more topical, partly because of advances in comparative biology and genetics, which allow old questions to be addressed with more resolution as well as the formulation of new questions (Coyne and Orr, 2004; SPECIATION-Network, 2011; Wolf et al., 2010). Nevertheless, the major questions in speciation biology are familiar; what are the relative roles of selection, drift and gene flow in promoting or retarding divergence? Selection can arise due to ecological adaptation and the role of natural selection or ecological specialisation in reducing gene flow between taxa has been highlighted in many recent reviews (Nosil et al., 2009; Sobel et al., 2010). Selection can also arise due to variation in mating success and sexual selection has been suggested to be a particularly important cause of speciation in animals (e.g. Panhuis et al., 2001). Nevertheless, finding strong evidence to support a particular role of sexual selection in speciation has proven to be far from straightforward, and some authors have questioned its importance (Ritchie, 2007; Kraaijeveld et al., 2010; Maan and Seehausen, 2011).

At one level a potential link between sexual selection and speciation seems obvious; many animals seem to differ primarily in sexually dimorphic traits such as the plumage of many birds, antlers of deer, beetles or trilobites, or behaviours such as courtship song or other signals. Comparative genomics is showing that sexual dimorphism extends to the genome, with genes that show sexually dimorphic expression evolving more quickly and displaying stronger signals of adaptive divergence (Ellegren and Parsch, 2007). But these apparent associations may be misleading, as many are largely anecdotal and non-quantitative. Sexual dimorphism is often thought to be a key signature of sexual selection and if dimorphism is obviously related to a trait primarily involved in sexual communication, the link may be more

Received Nov. 1, 2011; accepted Jan. 27, 2012.

∗ Corresponding author. E-mail: mgr@st-andrews.ac.uk

safely established. However, natural selection may act to promote or constrain divergence on dimorphic traits, potentially making such measures misleading. Further, ecological specialisation is often not studied where there is elaborate sexual dimorphism, and niche partitioning may occur between sexes as well as species. While numerous theoretical studies demonstrate that sexual signals and mate preferences rapidly evolve and could cause sexual isolation (e.g. Lande, 1981; Higashi et al., 1999; Kirkpatrick and Lavigne, 2002), it has been argued that sympatric species that only differ in behaviour are unlikely to persist without niche specialisation (Sobel et al., 2010). More quantitative studies are required to assess in detail the rates of evolution of ecological and sexual isolation, and both acting in concert may be a much more powerful force than either acting alone, especially if natural selection also acts on traits involved in mating behaviour. Comparative analyses suggest that reproductive isolation is greater where species show ecological differentiation (Funk et al., 2006).

The most common methodology adopted when asking if sexual selection promotes speciation is comparative analyses of rates of speciation in organisms differing in inferred levels of sexual selection. Sexual dimorphism, especially in plumage colouration in birds, was a major focus of early studies, which found a positive correlation between plumage dimorphism and species richness (Barraclough et al., 1995; Møller and Cuervo, 1998). Curiously, Owens et al. (1999) found an effect of plumage dimorphism but no effect of polyandry, presumably a more direct measure of the levels of sexual selection. In lizards, speciation rate is increased in sexually dichromatic lineages but reduced in lineages with greater size dimorphism (Stuart-Fox and Owens, 2003). In insects, some studies find a greatly increased rate of divergence in more polyandrous clades (Arnqvist et al., 2000) but arguably the most detailed studies across organisms failed to find any association between species richness and either size dimorphism or measures of polyandry (Gage et al., 2002; Morrow et al., 2003). Most recently, a meta-analysis of the comparative evidence found modest support for a role of sexual selection in speciation (Kraaijeveld et al., 2010). Interestingly, the strength of an association varied between study organisms (with fish having the strongest effect, though there were few studies) and between the measures used to infer the extent of sexual selection; sexual dichromatism had the most pronounced effect, but size dimorphism had an overall effect size of zero. Clearly there are many potential confounding variables in broad comparative studies, including ascertainment biases (are we more likely to define species in elaborately dichromatic lineages?), the reliability of relatively simple indices of the strength of sexual selection, and confounding factors such as the probability of extinction, which may be increased for more sexually selected species (Morrow and Pitcher, 2003).

Less intensively used in this area are studies of phylogeography and direct assessment of predictors of gene flow in the field. Boul et al. (2007) demonstrated less gene flow at neutral loci between populations of the Amazonian frog *Physalaemus petersi*, which differ in calls and preferences. Calls seem to predict species in some closely related groups of insects (Henry and Wells, 1990, Mendelson and Shaw, 2005) but do not always accurately correlate with genetic differences or patterns of reproductive isolation (Gleason and Ritchie, 1998). It seems unfortunate that there are fewer studies of this type; correlating gene flow with traits would seem a more direct method of assessing the importance of traits to divergence than broad comparative studies.

There has been one broad comparative study of the role of sexual selection in speciation in fish. Mank (2007) that sexually selected traits (mainly dichromatism or exaggerated male fin shape) are associated found across ray-finned fish with a higher rate of divergence. Similarly, closely related species flocks of cichlid fish in Nicaragua lakes are genetically structured more by nuptial colour than by trophic or niche diversification (Wilson et al., 2000). We have been studying speciation within one family of fish, the Goodeinae, live-bearing freshwater fish, endemic to Mexico. These show extreme variation in sexual dimorphism related to sexual selection, with some extremely dimorphic species and others virtually sexually monomorphic (Ritchie et al., 2005). Copulation requires female cooperation, and the range of sexual ornamentation in the males seems to explain the variation in morphological sexual dimorphism, at least anecdotally. We have previously completed a comparative study across all the species and a directed comparison of population differentiation between contrasting species. Here we concentrate on a new study of morphological variation amongst populations of one of the most sexually dimorphic species of the family *Girardinichthys multiradiatus*.

1 Materials and Methods

1.1 The studied species

The Goodeinae consist of around 34 extant species of fish, mainly from shallow fresh water drainages of cen-

tral Mexico. The group radiated in the Miocene and occupies the basins to the north of the Mexican neovolcanic belt. These are small (usually < 10 cm in length) omnivorous topminnows, except for a large (ca. 20 cm) piscivorous specialist (*Alloophorus robustus*), and a few species with a degree of specialization such as *Ameca splendens* (a plant feeder), the benthivorous *Zoogoneticus* spp. and the riverine genera *Ilyodon*, *Xenotaenia* and *Allodontichthys* (also a benthivore) (see Miller et al., 2005).

Molecular (mtDNA) phylogenies of the group (Doadrio and Domínguez, 2004; Webb et al., 2004) confirm their monophyly and that the group is around 15 million years old, suggesting their radiation was related to the complex volcanic history of the region. This radiation was predominantly allopatric, but ranges have subsequently changed dramatically due to volcanism (Webb, et al., 2004) allowing interactions between divergent species.

Previously, we quantified morphological measures of sexual dimorphism across wild caught samples of 25 species of the group and tested for an association between sexual dimorphism and time to speciation within the clade using comparative approaches. Based on reconstructed levels of sexual dimorphism, time till speciation was shorter in more dimorphic lineages in 10 vs. 6 comparisons, a non-significant difference (Ritchie, et al., 2005). In a different study (Ritchie, et al., 2007) we examined gene flow between populations within species. Our logic was that if gene flow was influenced by sexual selection, the drift-selection balance would differ between populations of species with contrasting levels of sexual selection. We compared two pairs of species with relatively high and low levels of sexual dimorphism. F_{ST} was higher for a given geographic distance (approximately 0.26 vs. 0.16 adjusted for distance) between both of the more dimorphic species. Furthermore, there was evidence of sex-biased gene flow between populations of the dimorphic species, consistent with female preference against immigrant males (Ritchie et al., 2007). Hence our studies were consistent with a potential influence of sexual selection, as evidenced by sexual dimorphism, being correlated with gene flow at the level of paired comparisons, although there was not a strong signal of sexual selection influencing speciation at the level of the family.

Girardinichthys multiradiatus, or Amarillo fish, is amongst the most sexually dimorphic goodeid species. Males have much larger and colourful median fins than females (Fig. 1), and perform complicated courtship displays (Zuarth and Macías Garcia, 2006; Zuarth et al., 2011). Studies of the behavioural ecology of this species suggest that sexual selection must be particularly important. Fin morphology influences female mate choice; females preferentially associate with males possessing larger fins (Macías Garcia et al., 1994), and males with reduced sexual dimorphism in body shape are discriminated against (Arellano-Aguilar and Macías Garcia, 2008). Fin morphology also increases predation risk (Macías Garcia et al., 1994; Macías Garcia et al., 1998). Colour, including UV, is also a criterion for female mate choice (Macías Garcia and Burt de Perera, 2002). Ornaments in the Amarillo are condition-dependent, because embryonic exposure to pesticides leads to suboptimal expression of sexually dimorphic fins, colour, and courtship displays (Arellano-Aguilar and Macías Garcia, 2008).

Because male morphology is heritable (Arellano-Aguilar and Macías Garcia, 2008), influences reproductive success via female mate choice, and is opposed by local natural selection (Macías Garcia et al 1998) the species appears to be a good candidate with which to explore the possible link between sexual selection, sexual dimorphism and gene flow at the level of a phylogeographic analysis. If sexual dimorphism evolves as

Fig. 1 Male (above) *G. multiradiatus* have larger and more colourful fins than females (below: this is a pregnant female)

Forced inseminations cannot occur in this family of fish, since sperm transfer is only aided by the small spermatopodium (at the front of the male's anal fin), which is not an intromittent copulatory organ.

a response to sexual selection and divergence influences gene flow, we predicted that a) sexual dimorphism would be variable between populations b) males may differ more between populations than females (if female preferences have selected for exaggeration of male morphology, including ornaments) c) either sexual dimorphism or male morphology would predict patterns of behavioural sexual isolation between populations d) genetic differentiation between populations would be correlated with sexual dimorphism or behavioural sexual isolation between populations. Here we test each of these predictions.

1.2 Sampling and morphology

Girardinichthys multiradiatus were sampled at several localities in Central Mexico, which encompass the majority of its geographic range (Gesundheit and Macías Garcia 2005; Fig. 2). These include ponds, reservoirs and rivers in three major catchments, although the bulk of the populations are found in the upper Lerma River basin. Topography suggests that the few localities occupied by the Amarillo in the Balsas basin were colonised independently, whereas the small area of the Panuco where this species is found was probably reached through a single region lying adjacent to one of the lowest points of the watershed divide. Collections were conducted with 3m seine nets hauled towards the shores (where adults congregate; Macías Garcia et al., 1994). Fish were anaesthetised and photographed upon capture, when a small portion of the tail fin was collected for DNA extraction. Fish were the given commercial anti-stress treatment before being transported to the laboratory, where they were housed in 40L tanks provided with filters and aeration (see González Zuarth and Macías Garcia, 2006).

Morphology was examined from digital photographs taken on anaesthetised fish; as reported in González-Zuarth and Macías Garcia (2006). We included six measures that have been implicated in female mate choice (standard length, body depth, and length and breadth of the dorsal and anal fins), and four that have not (eye and pupil diameter, depth of the caudal peduncle and length of the tail fin). Morphological variables were standardised to a mean of zero and a standard deviation of one then subject to a canonical discriminant

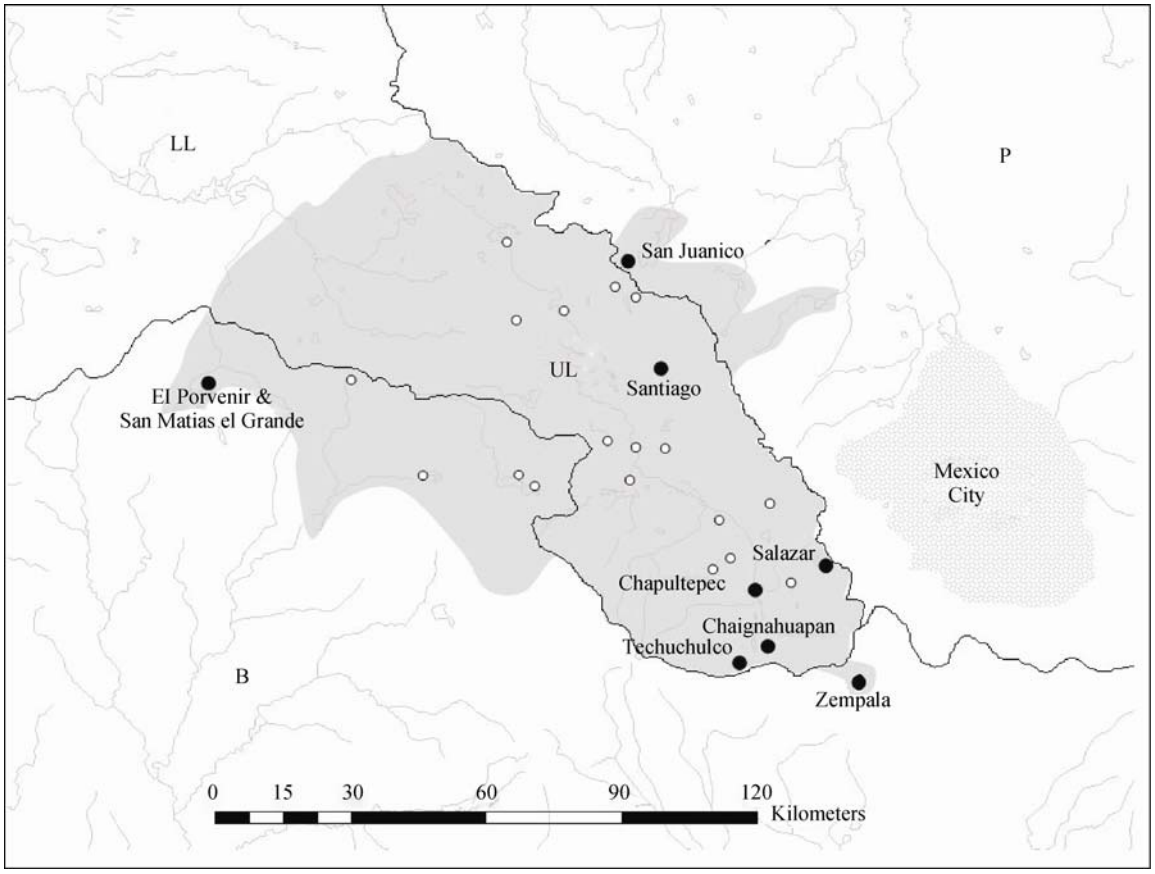


Fig. 2 Hydrographical map of Central Mexico

Black lines run along the watershed divides between the Panuco (P), Balsas (B) and Lerma (upper Lerma, UL; lower Lerma, LL) basins. Circles indicate localities containing substantial populations of *G. multiradiatus* known to us; larger, solid circles show the locations we have studied in more detail.

analysis (with population and sex giving 14 discriminant groups, i.e. population 1 males, population 1 females, population 2 males, and so on for all populations). This was in order to 1) quantify global sexual dimorphism as the canonical distance between males and females from each population, 2) quantify differences in sexual dimorphism between populations, as the pairwise population comparisons of sexual dimorphism, 3) assess whether male morphology diverges more rapidly than female morphology, and 4) assess whether the differences in male or female morphology between populations are related to the amount of genetic and/or behavioural divergence (assortative mating). We did not correct for size variation before completing the multivariate analysis because differences in size can be an important component of sexual dimorphism, and is one of the most widely used proxies for sexual selection (e.g. Gage et al., 2002; Kraaijeveld et al., 2010). However, our conclusions remain unchanged if size is not included in the analysis.

1.3 Behavioural analyses

Data on behavioural isolation and male and female morphologies were collected from seven populations, including those described by González Zuarth and Macías Garcia (2006), with the addition of Techuchulco and Chignahuapan. Briefly, each of between 30 and 35 females from each population was exposed to a pair of males; one of her own population and another from a different population. Every female was presented a different pair of males (which were not re-used), and we performed 5–7 replicates of each combination. The mean of these replicates was used in the analysis. Males were introduced into the observation tank within plastic bags containing water from their home tanks, thus avoiding the stress associated with netting and water changes. Presentation of males within plastic bags allows the females to approach each male from a variety of directions as they do in the field, and also facilitates the occurrence of copulation attempts (copulatory embraces), which do not occur across rigid walls. Males were located at opposite ends of the tank and after 10 min of habituation the female was introduced and her behaviour recorded. We quantified the number and duration of visits (approaches to within one female body length) to each male (previously the frequency of copulation attempts was used to demonstrate that the duration of visits is a good predictor of willingness to mate, which translates into paternity; see González Zuarth and Macías Garcia, 2006; Macías Garcia and Saborío, 2004). Since females from each of the seven populations were tested with six possible combinations of sympatric - allopatric males, we obtained 42 scores of pre-mating isolation, which were calculated as the difference between the mean time (in seconds) spent with sympatric versus allopatric males. We then added the scores of reciprocal tests (i.e. females of population A choosing between males from population A and B, and females of population B choosing between males of the same two localities). This generated 21 measures of the strength of pre-mating isolation ranging from -67 (negative values indicating that females of at least one of the two localities discriminate against their own males) to 142 (mean = 34.5 ± 12.8 *SE*). These 21 scores satisfied assumptions of normality and thus were not transformed before the analyses.

1.4 Genotyping

DNA was extracted from fin clips of individuals from ten populations of *G. multiradiatus* (Fig. 2) using the Purgene DNA isolation method. Primers for microsatellite loci have been previously designed for various fish species and were obtained for this study from Boto and Doadrio (2003) and Hamill et al. (2007). In total fifteen loci were surveyed using primers designed for the Goodeinae, these were: XC18, XC25, AS2, AS4, AS5, CA6, CA8, CA10 CA12, IW193 IW196 developed by Hamill et al.(2007) and ZT1.3, ZT1.6, ZT1.7and ZT1.9 by Boto and Doadrio (2003). Hamill et al. (2007) found only 7 of these loci to be polymorphic in *G. multiradiatus* with three displaying little variation. Here we tested all fifteen loci and found only five (XC18, IW196, ZT1.6, ZT1.9 and CA12) polymorphic enough for population genetic analyses. In total 408 individuals from the 10 populations were analysed for five microsatellite loci (see Table 4a for numbers per population). One primer for each locus was fluorescently tagged and PCR was carried out in multiplexed groups using the Qiagen Multiplex PCR Kit and genotyped on a Beckman Coulter CEQ 8000XL.

1.5 Population genetic analyses

Micro-Checker (Van Oosterhout et al., 2004) was used to check the reliability of the microsatellite genotype data and test for Hardy-Weinberg Equilibrium (HWE). Tests of HWE for each population and each locus, linkage disequilibrium, population differentiation (F-statistics) as calculated by Weir and Cockerham (1984) and migration estimates (N_m) were calculated with GenePop (Raymond and Rousset, 1995; Rousset, 2008) and Arlequin version 3.1 (Excoffier et al., 2005) using Fisher's exact tests (Guo and Thompson, 1992) with Bonferroni correction when multiple tests were

performed.

Metrics of genetic differentiation, such as F_{ST} and G_{ST}, have been criticized as inaccurate measures of population differentiation when gene diversity is high because they approach zero, even when subpopulations are highly structured (Jost, 2008). Jost (2008) suggested an estimator of actual differentiation, D_{est} that takes into account gene diversity, which we calculated using the online program SMOGD (Crawford, 2010). An isolation-by-distance (IBD) analysis was performed using the isolation-by-distance service (IBDWS) version 3.15 (Jensen et al., 2005) through a Mantel test with 30,000 permutations and custom software. Due to our sampling both within and between watersheds we examined IBD over the entire dataset as well as looking at only within the upper Lerma River basin. Finally, matrices of simple linear geographic distance), and our measures of sexual isolation and sexual dimorphism between the seven populations in common were constructed. A partial mantel test was used to examine the relationships between genetic distance (D_{est}) and geographic distance, sexual isolation and sexual dimorphism. A sex biased dispersal analysis was carried following Goudet et al. (2002) as implemented in FSTAT version 2.9.3 (Goudet, 2001).

2 Results

2.1 Evolution of sex dimorphism

Sexual dimorphism, calculated as the between sex differences for each population in multivariate measures of morphological variation, differed between populations (range 7.2–8.5). Standard length and dorsal and caudal fin size had the largest loadings on the canonical variates (Table 1), so both size dimorphism and relative fin size contribute to morphological divergence between sexes and populations. Measures derived from males and females were highly correlated between the sexes (Fig. 3) but values for males were higher. In order to assess whether the rate at which populations differ in morphology between males and females differs significantly, we compared a) the mean between-population canonical distances of males with that of females (males; 2.79±1.8; females 0.92 ± 0.43; t_{paired} = 5.8, n=21, P<0.0001; Fig. 3 [also, the slope is significantly >1]; Table 2), and b) the scores from the first two canonical variables between sexes and populations (Table 3). There were significant differences both between sexes (CV1) and between populations (CV1 and CV2), and the interaction was also significant for both canonical variables, confirming that the extent of differences were not the same between sexes.

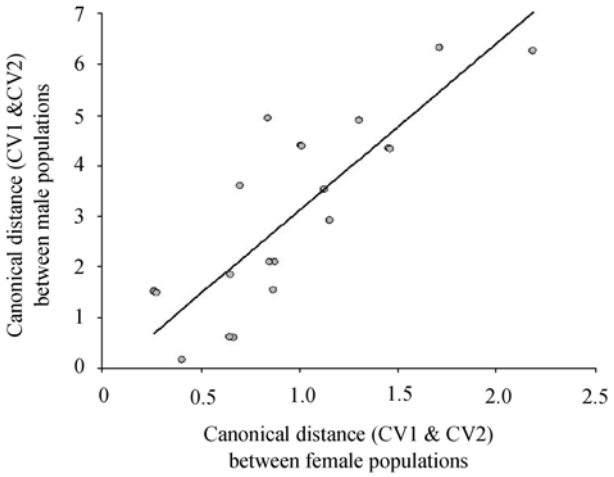


Fig. 3 Canonical distances from the first two canonical variates (92.5% of variance explained; Table 1) of a discriminant analysis of morphological variation amongst populations and sexes

Table 1 Standardised canonical coefficients from the first two canonical variates of a discriminant analysis to calculate sexual dimorphism and morphological divergence between populations of males and of females

Variable	Canonical Variate	
	CV1	CV2
Standard length	**1.024009**	**1.323412**
Length of anal fin	–0.020288	–0.841586
Length of caudal fin	0.188106	**–1.355198**
Length of dorsal fin	–0.689712	**2.128972**
Breadth of anal fin	–0.473283	–0.461595
Breadth of dorsal fin	**–1.005069**	**–1.164784**
Eye diameter	0.113949	0.311531
Pupil diameter	0.175604	–0.36217
Body depth	–0.072345	0.155999
Depth of caudal peduncle	0.619857	0.401072
Eigenvalue	16.10705	2.400353
% of variance explained	80.5	12
Cumulative %	80.5	92.5
F-value	11.4	5.4
Numerator DF	130	108
Denominatior DF	1576	1439.4
P	0	0
Wilk's Lambda	0.004907	0.083937

Table 2 Classification of males and females per population of origin based on the canonical analysis reported in Table 1

		Belonged to													
		♀							♂						
	Classified in	**Ch**	**P**	**Sal**	**Sgo**	**SJ**	**T**	**Z**	**Ch**	**P**	**Sal**	**Sgo**	**SJ**	**T**	**Z**
	Chignahuapan	**9**	0	0	0	0	2	0	0	0	0	0	0	0	0
	Porvenir	1	**13**	1	1	1	0	1	0	0	0	0	0	0	0
	Salazar	0	2	**4**	5	3	0	3	0	0	0	0	0	0	0
♀	**S**antia**go**	0	1	3	**5**	1	1	3	0	0	0	0	0	0	0
	San **J**uanico	1	0	2	2	**10**	0	2	0	0	0	0	0	0	0
	Techuchulco	4	0	0	2	1	**12**	0	0	0	0	0	0	0	0
	Zempoala	0	0	0	1	2	0	**3**	0	0	0	0	0	0	0
	Chignahuapan	0	0	0	0	0	0	0	**10**	0	0	1	0	5	0
	Porvenir	0	0	0	0	0	0	0	0	**15**	0	4	1	0	2
	Salazar	0	0	0	0	0	0	0	0	0	**10**	2	0	0	1
♂	**S**antia**go**	0	0	0	0	0	0	0	0	1	0	**11**	1	0	3
	San **J**uanico	0	0	0	0	0	0	0	0	3	4	4	**8**	0	0
	Techuchulco	0	0	0	0	0	0	0	2	0	0	0	0	**10**	0
	Zempoala	0	0	0	0	0	0	0	0	0	1	10	2	0	**5**
	Total	15	16	10	16	18	15	12	12	19	15	32	12	15	11
	Proportion correct	0.6	0.8	0.4	0.3	0.56	0.8	0.25	0.8	0.8	0.67	0.3	0.67	0.67	0.45

Table 3 ANOVAs of CV1 and CV2 (from Table 1) examining the consistency of sex differences amongst populations

Analysis of Variance Table, CV1			
Source Term	***df***	***F* ratio**	***P* level**
Sex	1	686.21	<0.0001*
Population	6	3.38	0.0033*
Sex*Population	6	4.32	0.0004*
Error	204		
Total (Adjusted)	217		
Total	218		
Analysis of Variance Table, CV2			
Sex	1	0.59	0.473
Population	6	62.34	<0.0001*
Sex*Population	6	19.19	<0.0001*
Error	204		
Total (Adjusted)	217		
Total	218		

2.2 Sexual isolation and sexual dimorphism

To explore the relationship between sexual isolation and the degree of morphological differentiation we regressed our measure of assortative mating on the population differences in sexual dimorphism and on the canonical distances between male morphology across populations. Differences in sexual dimorphism were unrelated to pre-mating isolation (Fig. 4a Mantel P=0.2), whereas, curiously, the magnitude of the differences in male morphology between populations was negatively correlated to the degree of pre-mating isolation (i.e. the greater the difference in male morphology, the lower the pre-mating barriers; Fig. 4b; Mantel P=0.03).

2.3 Genetic differentiation

The five polymorphic microsatellite loci surveyed were found to have an allelic diversity ranging from nine to 35 alleles and an observed heterozygosity of 0.038–0.921 (Table 4b). Locus CA12 had a large number of alleles and demonstrated significant deviation from HWE in half the populations (Table 5), however results were comparable with or without this locus. No other loci demonstrated consistent significant deviations from HWE overall, though 4 of the 5 loci showed significant heterozygote deficit in the Chignahuapan and San Pedro Techuchulco populations. Population substructure is possible for the Chignahuapan population as individuals came from four geographically close sampling localities. Possible null alleles were detected by Micro-Checker in only three populations for locus CA12, two populations for locus XC18 and one for loci IW196 and ZT1.9, with no evidence of large allele

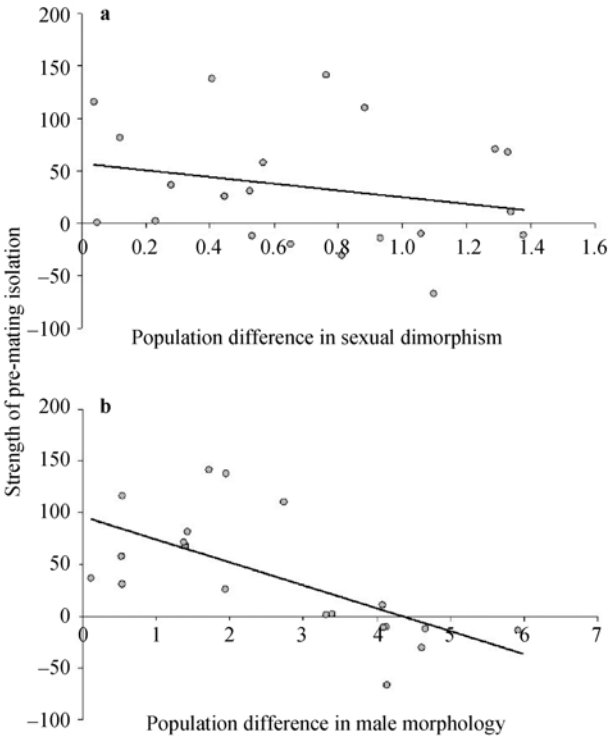


Fig. 4 The strength of pre-mating isolation against population differences in sexual dimorphism (a) and differences in male morphology (b)

dropout or scoring errors due to stuttering in any population.

Estimates of F-statistic genetic differentiation (F_{ST}) for each population comparison were significant (after Bonferroni correction) except for that between El Porvenir and San Matias el Grande. Measures of D_{est} were taken as the harmonic mean across all loci for each population pair and ranged from 0.011 to 0.598 with El Porvenir and San Matias el Grande populations being highly differentiated from all else, but not each other (Table 6). The Zempoala population was also found to highly differentiated from all groups except the Salazar lake population (Table 6).

Comparison of the genetic and trait data showed that morphological differentiation between populations appeared arbitrary, in the sense that it is not significantly related to genetic distance in either sex (Fig. 5: Mantel P=0.85 for males, 0.29 for females). Results of isolation-by-distance analyses showed a significant correlation of genetic distance (D_{est}) to geographic distance (Mantel P=0.0047; Fig. 6). However, this is clearly superimposed on a larger difference between populations from different watersheds (Fig. 2). IBD within the Lerma basin alone was non-significant (Mantel P=0.46).

Table 4 Summary of microsatellite sampling site locations, including sample size per population (*n*) (a), microsatellite variability in *G. multiradiatus* (b)

(a)

Population	Location	*n*
Lake of Chinahuapan	19 08' N, 99 29' W	94
Chapultepec	19 13' N, 99 33' W	30
El Porvenir	19 40' N, 100 38' W	30
Salazar	19 17' N, 99 24' W	28
Santiago	19 44' N, 99 41' W	44
San Juanico	19 55' N, 99 43' W	42
San Matias el Grande	19 42' N, 100 42' W	29
San Pedro Techuchulco	19 06' N, 99 31' W	37
Temascalcingo	19 55' N, 100 0' W	30
Zempoala	19 19' N, 99 19' W	44

(b)

Locus	Size range	Number of Alleles
XC18	248-298	22
ZT1.6	232-250	9
IW196	191-221	9
ZT1.9	357-429	17
CA12	127-243	35

Partial Mantel tests comparing genetic distance to sexual isolation and sexual dimorphism were also non-significant (P=0.35 and P=0.1 respectively). Further, a pattern of sex-biased dispersal was not seen in our data (P>0.45). Thus, the only variable that significantly explains genetic differentiation of *G. multiradiatus* populations is geographic distance, in a manner compatible with isolation across watersheds.

3 Discussion

Testing for a potential role of sexual selection in speciation and population divergence has been undertaken using a variety of approaches and at a number of scales, from broad comparative studies to analysis of gene flow in individual species. Results provide at best limited support for a clear role of sexual selection in speciation (Ritchie 2007; Kraaijeveld et al., 2010). This may be because the importance of sexual selection in causing reproductive isolation has been overestimated and natural selection may be more effective (Sobel et al., 2010). Mating signals may diverge quickly only because of an interaction with ecological selection, if they allow assortative mating between entities under divergent

Table 5 Summary of microsatellite data by population and locus

Population	Parameter	Loci				
		XC18	*ZT1.6*	*IW196*	*ZT1.9*	*CA12*
Chignahuapan	Genotypes	68	75	90	77	68
	H_O	0.706*	0.360	0.122*	0.766*	0.676*
	H_E	0.921	0.422	0.159	0.879	0.91852
Chapultepec	Genotypes	23	28	28	27	28
	H_O	0.957	0.286	0.036	0.704	0.821
	H_E	0.927	0.371	0.036	0.774	0.901
El Porvenir	Genotypes	-	-	-	27	26
	H_O	-	-	-	0.667	0.385
	H_E	-	-	-	0.488	0.384
Salazar	Genotypes	28	27	27	28	28
	H_O	0.821	0.333	0.370	0.321	0.607*
	H_E	0.767	0.372	0.391	0.456	0.756
Santiago	Genotypes	44	43	43	44	43
	H_O	0.841	0.418*	0.163	0.795	0.535*
	H_E	0.872	0.476	0.153	0.827	0.854
San Juanico	Genotypes	38	34	-	39	35
	H_O	0.763	0.265	-	0.769	0.714*
	H_E	0.838	0.237	-	0.712	0.865
San Matias El Grande	Genotypes	-	28	-	28	29
	H_O	-	0.107	-	0.678*	0.345
	H_E	-	0.103	-	0.477	0.373
San Pedro Techuchulco	Genotypes	26	23	30	23	21
	H_O	0.461*	0.565*	0.166	0.478*	0.619*
	H_E	0.782	0.674	0.220	0.797	0.941
Temascalcingo	Genotypes	29	18	26	29	26
	H_O	0.828	0.333	0.038*	0.828	0.769
	H_E	0.899	0.390	0.112	0.802	0.916
Zempoala	Genotypes	43	35	43	42	38
	H_O	0.581	0.400	0.465	0.595	0.605
	H_E	0.621	0.487	0.444	0.657	0.722

Genotypes: number of genotypes. H_O: observed heterozygotsity. H_E: expected heterozygosity. - : locus monomorphic in that population.
*: significant deviation from Hardy-Weinberg equilibrium (P<0.05).

Table 6 Pairwise D_{est} estimates for each population pair

	Lake of Chinahuapan	Chapultepec	El Porvenir	Salazar	Santiago	San Juanico	San Matias el Grande	San Pedro Techuchulco	Temascalcingo	Zempoala
Lake of Chinahuapan	--	0.0474	0.4258	0.1878	0.0425	0.1089	0.4297	0.1086	0.0276	0.2948
Chapultepec	--	--	0.3183	0.2509	0.0763	0.0796	0.32	0.1154	0.0668	0.3237
El Porvenir	--	--	--	0.2424	0.429	0.3322	–0.001	0.3129	0.4303	0.4019
Salazar	--	--	--	--	0.219	0.2129	0.2529	0.2086	0.1681	0.0659
Santiago	--	--	--	--	--	0.1067	0.4287	0.118	0.0351	0.3405
San Juanico	--	--	--	--	--	--	0.3322	0.1452	0.0875	0.3034
San Matias el Grande	--	--	--	--	--	--	--	0.309	0.4304	0.4056
San Pedro Techuchulco	--	--	--	--	--	--	--	--	0.101	0.2521
Temascalcingo	--	--	--	--	--	--	--	--	--	0.2916
Zempoala	--	--	--	--	--	--	--	--	--	--

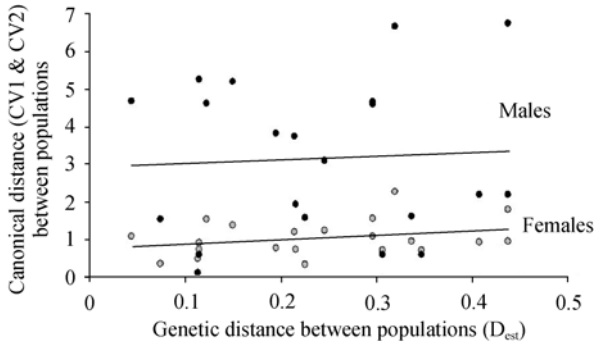


Fig. 5 Morphological divergence between populations of males (filled circles) and females (shaded circles) against genetic distance

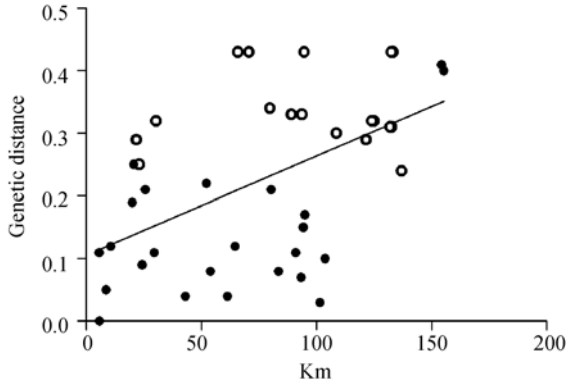


Fig. 6 Isolation by distance

Genetic distance (D_{est}) plotted against geographic distance (Km). Open circles represent between watershed comparisons and filled circles within watershed comparisons.

ecological selection (Van Doorn et al., 2009). Sexual selection may actively oppose speciation under some conditions, for example if male competition is a more powerful force than assortative female preferences (Parker and Partridge, 1998). Alternatively, a problem that has been highlighted with comparative studies of sexual selection and speciation is that the proxies used to estimate variation in the intensity of sexual selection may be misleading (Gonzalez-Voyer and Kolm, 2011; Kraaijeveld et al., 2010). Our study has allowed a detailed examination of the evolution of divergence in sexual dimorphism and assortative mating, yet neither has shown any correlation with gene flow. Our prediction was that greater sexual dimorphism would evolve in populations where female preferences were stronger and that such populations would also build up genetic differences due to more restrictive matings (Ritchie et al., 2007).

We quantified sexual dimorphism among multiple populations. This dimorphism was due to both sexual size differences, but also differences in fin morphology. Size dimorphism could occur for a number of reasons, including sexual selection, but differences in fin shape seem almost certainly to be influenced by sexual selection. Fins are used extensively during courtship interactions and the exaggerated fin shape (and colour) of male fins in this species suggests they at least partly function as an 'ornament', i.e. a target of female mating preferences. We therefore expected to find an association between our measure of mating isolation and either the magnitude of the morphological divergence between populations (i.e. differences in sexual dimorphism or in male morphology) or the degree of genetic divergence. In general, local divergence of male mating traits is expected to arise due to 1) differences in magnitude of an ornament that can be expressed given the local ecology, and 2) differences in the magnitude of an ornament preferred by the majority of the local females. Indeed, from classical Fisherian models, it is expected that populations are driven apart by female mate choice only to the point where they are counter-selected by natural selection, where they are expected to remain. It is possible that variation in sexual dimorphism reflects ecological variation in counter-selection rather than variation in female preferences. *Girardinichthys multiradiatus* is an endangered fish with their natural habitat being lost (for example, since we began our studies on this species several populations have disappeared through desiccation, e.g. at Ignacio Ramirez, Ixtlahuaca and Acambay) rapidly through increased land use, tourism and pollution (De La Vega et al., 1997). Pollutants such as methyl parathion can affect the expression of male traits and influence male attractiveness to females (Arellano-Aguilar and Macías Garcia, 2008). Recent change in selection pressures, such as pollution, may have led to changes in the interplay of natural and sexual selection and thus create a disjunction between more historical genetic variation and our proxies for sexual selection. Further, invasive species such as the poeciliid *Heterandria bimaculata* (Ramírez Carillo and Macías Garcia, 2012) have played a big part in influencing the conservation status of *G. multiradiatus* through reproductive interference and conceivably may have caused a more recent shift in mate choice than the genetic data could detect.

It is also possible that the expectation of different levels of sexual selection and assortative mating may not be met in many systems. All populations could have directional female preferences for exaggerated fins, so it is not necessarily appropriate to translate differences in sexual ornamentation to potential assortative mating

(Price, 1998). There are very few systems where female preference functions amongst populations have been examined and then compared to variation in male traits, and those that have do not always show the covariance that would be predicted by correlated coevolution (Prum, 2010), which is necessary to underpin assortative mating. The fact that we see male morphology diverge between populations more quickly than female morphology supports the assumption that male morphology is under stronger selection than female morphology, however our expected link between morphological divergence and assortative mating is not supported. Indeed, the variation in sexual isolation we find is in fact negatively correlated with divergence in male morphology and sexual dimorphism. If females all have open-ended preferences for extreme males, they could show disassortative preferences when paired with males from more dimorphic populations. Perhaps our results also suggest more complicated preferences such as an attraction to novel male morphologies, which would more explicitly predict disassortative mating.

Although our measure of genetic divergence is derived from relatively few microsatellite markers we do detect isolation by distance and an analysis of mtDNA sequence variation showed a similar pattern, though there was limited variation (unpublished data). There is no suggestion that gene flow is related to either morphological variation or potential assortative mating. This is surprising as a previous study of F_{ST} amongst populations of four species of Goodeids did suggest that genetic differences between populations were greater in more dimorphic species. We believed the pattern would be greater in *G. multiradiatus*, as this is among the most sexually dimorphic species of the group. The first and most obvious explanation for these results is that sexual selection does not influence population divergence in this species. However, it is also possible that we did not assess a trait that predicts assortative mating. The exaggerated fins of male *G. multiradiatus* are highly colourful, and we have not assessed colour variation. Also, there is evidence in birds and some fish to suggest that colour patterns in the UV range of the light spectrum can be detected by females and can influence mate choice (Bennett and Cuthill, 1994; Smith et al., 2002). Further, we did not allow courting couples to have physical contact thus discounting the use of pheromones in mate choice. Other possible traits might mediate mate choice and thus be under sexual selection and more detailed analyses of courtship behaviour could be more revealing, as subtle changes in the behavioural courtship elements may be involved. However, the steady pattern of isolation by distance superimposed on divergence between watersheds is simply most compatible with genetic differences at these neutral markers building up by genetic drift despite the obvious (and quite striking) variation in sexual morphologies and behaviour in this sexually dimorphic species.

In general, our studies exploring a potential link between sexual dimorphism, behaviour and evolutionary divergence in the Goodeid fish point to the contrasting results seen at different levels of analysis. They highlight that simple measures of predictors of the extent of sexual selection amongst species or populations may be misleading. While some trait comparisons may predict gene flow in one context, the complex dynamics of sexual selection may mean that correlations are not maintained across different comparisons or scales of analysis (in their review of comparative evidence, Kraaijeveld et al., 2010, found that dichromatism and mating system had significantly greater mean effect sizes for an association with speciation than simple size dimorphism). Variation in a frog call or an elaborate fin may predict gene flow between certain populations, but why should this be generally true across multiple species? Furthermore, the divergence in sexually selected male traits may not accurately predict assortative mating, which depends on the interaction between variation in male traits and female preferences, both of which may be under multiple sources of selection.

Acknowledgments We would like to thank Martina Di Fonzo and Jenny Wright for their work in the lab and Edgar Avila who participated in fish collection, husbandry and behavioural assays, and three reviewers for helpful comments. Fieldwork was supported with a grant from CONACyT to CMG and molecular work with NERC grant NE/E008216/1 to MGR & JAG.

References

Arellano-Aguilar O, Macías Garcia C, 2008. Exposure to pesticides impairs the expression of fish ornaments reducing the availability of attractive males. Proc. Roy. Soc. Lond. BBio. 275: 1343.

Arnqvist G, Edvarrdsson M, Friberg U, Nilsson T, 2000. Sexual conflict promotes speciation in insects. Proc. Natl. Acad. Sci. USA 97: 10460–10464.

Barraclough TG, Harvey PH, Nee S, 1995. Sexual selection and taxonomic diversity in passerine birds. Proc. Roy. Soc. Lond. B Bio. 259: .211–215.

Bennett A, Cuthill I, 1994. Ultraviolet vision in birds: What is its function? Vision Res. 34: 1471–1478.

Boto L, Doadrio I, 2003. Polymorphic microsatellites in two different species of the genus *Zoogoneticus* Meek, 1902 (Goodeidae, Actynopterygii). Mol. Ecol. Notes 3: 70–72.

Boul KE, Funk WC, Darst CR, Cannatella DC, Ryan MJ, 2007. Sexual selection drives speciation in an Amazonian frog. Proc. Roy. Soc. Lond. B Bio. 274: 399–406.

Coyne JA, Orr HA, 2004. Speciation. Sunderland, Mass: Sinauer.

Crawford NG, 2010. SMOGD: Software for the measurement of genetic diversity. Mol. Ecol. Resources 10: 556–557.

De La Vega SMY, Martinez TL, Macias GC, 1997. Bioaccumulation of methyl parathion and its toxicology in several species of the freshwater community in Ignacio Ramirez Dam in Mexico. Ecotox. Environ. Safe. 38: 53–62.

Doadrio I, Domínguez O, 2004. Phylogenetic relationships within the fish family Goodeidae based on cytochrome b sequence data. Mol. Phylogenet. Evol. 31: 416–430.

Ellegren H, Parsch J, 2007. The evolution of sex-biased genes and sex-biased gene expression. Nat. Rev. Gen. 8: 689–698.

Excoffier L, Laval G, Schneider S, 2005. Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47.

Funk DJ, Nosil P, Etges WJ, 2006. Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. Proc. Natl. Acad. Sci. USA 103: 3209–3213.

Gage MJG, Parker GA, Nylin S, Wiklund C, 2002. Sexual selection and speciation in mammals, butterflies and spiders. Proc. Roy. Soc. Lond. B Bio. 269: 2309–2316.

Gesundheit P, Macías Garcia C, 2005. Biogeografía cladística de la familia Goodeidae. In: Bousquets JL, Morrone JJ eds. Regionalización biogeográfica en Iberoamérica y tópicos afines: Primeras Jornadas Biogeográficas de la Red Iberoamericana de Biogeografía y Entomología Sistemática. UNAM: Mexico City, 319–338.

Gleason JM, Ritchie MG, 1998. Evolution of courtship song and reproductive isolation in the *Drosophila willistoni* species complex: Do sexual signals diverge the most quickly? Evolution52: 1493–1500.

González Zuarth C, Macías Garcia C, 2006. Phenotypic differentiation and pre-mating isolation between allopatric populations of *Girardinichthys multiradiatus*. Proc. Roy. Soc. Lond. B Bio. 273: 301–307.

Gonzalez Zuarth CA, Vallarino A, Macías Garcia C, 2011. Female responsiveness underlies the evolution of geographic variation in male courtship between allopatric populations of the fish *Girardinichthysmultiradiatus*. Evol. Ecol. 25: 831–843.

Gonzalez-Voyer A, Kolm N, 2011. Rates of phenotypic evolution of ecological characters and sexual traits during the Tanganyikan cichlid adaptive radiation. J. Evolution.Biol.24: 2378–2388.

Goudet J, 2001. FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9. 3). http://www2.unil.ch/popgen/ softwares/fstat.htm

Goudet J, Perrin N, Waser P, 2002. Tests for sex-biased dispersal using bi-parentally inherited genetic markers. Mol. Ecol. 11: 1103–1114.

Guo SW, Thompson EA, 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48: 361–372.

Hamill RM, Webb SA, Macías Garcia C, Graves JA, Magurran AE et al., 2007. Comparison of genetic diversity at microsatellite loci in near-extinct and non-endangered species of Mexican goodeine fishes and prediction of cross-amplification within the family. J. Fish Biol. 70 (Sup. A): 16–32.

Henry CS, Wells MM, 1990. Geographical variation in the song of *Chrysoperla plorabunda* (Neuroptera: Chrysopidae) in North America. Ann. Entomol. Soc. Am. 83: 317–325.

Higashi M, Takimoto G, Yamamura N, 1999. Sympatric speciation by sexual selection. Nature 402: 523–526.

Jensen J, Bohonak A, Kelley S, 2005. Isolation by distance, web service. BMC Genetics 6: 13.

Jost L, 2008. GST and its relatives do not measure differentiation. Mol. Ecol. 17: 4015–4026.

Kirkpatrick M, Ravigne V, 2002. Speciation by natural and sexual selection: Models and experiments. Am. Nat. 159: S22–S35.

Kraaijeveld K, Kraaijeveld-Smit FJL, Maan ME, 2010. Sexual selection and speciation: The comparative evidence revisited. Bio.Rev. 86: 367–377.

Lande R, 1981. Models of speciation by sexual selection on polygenic traits. Proc. Natl. Acad. Sci. USA78: 3721–3725.

Maan ME, Seehausen O, 2011. Ecology, sexual selection and speciation. Ecol. Lett. 14: 591–602.

Macías Garcia C, Burt de Perera T, 2002. Ultraviolet based female preferences in a viviparous fish. Behav. Ecol. Sociobiol. 52: 1–6.

Macías Garcia C, Jimenez G, Contreras B, 1994. Correlational evidence of a sexually-selected handicap. Behav. Ecol. Socoibiol. 35: 153–259.

Macías Garcia C, Saborío E, Berea C, 1998. Does male-biased predation lead to male scarcity in viviparous fish? J. Fish Biol. 53: 104–107.

Macías-Garcia C, Saborio E, 2004. Sperm competition in a viviparous fish. Environ. Biol. Fish. 70: 211–217.

Mank JE, 2007. Mating preferences, sexual selection and patterns of cladogenesis in ray-finned fishes. J. Evolution. Biol. 20: 597–602.

Mendelson TC, Shaw KL, 2005. Rapid speciation in an arthropod. Nature 433: 375–376.

Miller RR, Minckley WL, Norris SM, 2005. Freshwater Fishes of Mexico. Chicago, Ill: University of Chicago Press.

Møller AP, Cuervo JJ, 1998. Speciation and feather ornamentation in brids. Evolution 52: 859–869.

Morrow EH, Pitcher TE, 2003. Sexual selection and the risk of extinction in birds. Proc. Roy. Soc. Lond. B Bio. 270:

1793–1799.

Morrow EH, Pitcher TE, Arnqvist G, 2003. No evidence that sexual selection is an 'engine of speciation' in birds. Ecol. Lett. 6: 228–234.

Nosil P, Harmon LJ, Seehausen O, 2009. Ecological explanations for (incomplete) speciation. Trends Ecol. Evol. 24: 145–156.

Owens IPF, Bennett PM, Harvey PH, 1999. Species richness among birds: Body size, life history, sexual selection or ecology? Proc. Roy. Soc. Lond. B Bio. 266: 933–939.

Panhuis TM, Butlin R, Zuk M, Tregenza T, 2001. Sexual selection and speciation. Trends Ecol. Evol. 16: 364–371.

Parker GA, Partridge L, 1998. Sexual conflict and speciation. Philos. T. Roy. Soc. B 353: 261–274.

Price T, 1998. Sexual selection and natural selection in bird speciation. Philos. T. Roy. Soc. B 353: 251–260.

Prum RO, 2010. The Lande-Kirkpatrick mechanism is the null model of evolution by intersexual selection: Implications for meaning, honesty, and design in intersexual signals. Evolution 64: 3085–3100.

Ramírez Carillo E, Macías Garcia C, 2012. (in prep) Climate warming and invasions: A thermal niche displacement dilemma for native fish.

Raymond M, Rousset F, 1995. GENEPOP (version 1.2): Population genetics software for exact tests and ecumenicism. Journal of Heredity. 86: 248.

Ritchie MG, 2007. Sexual selection and speciation. Annual Review of Ecology Evolution and Systematics. 38: 79–102.

Ritchie MG, Hamill RM, Graves JA, Magurran AE, Webb SA et al., 2007. Sex and differentiation: Population genetic divergence and sexual dimorphism in Mexican goodeid fish. J.Evol.Biol. 20: 2048–2055.

Ritchie MG, Webb SA, Graves JA, Magurran AE, Macias-Garcia C, 2005. Patterns of speciation in endemic Mexican Goodeid fish: Sexual conflict or early radiation? J. Evol. Biol. 18: 922–929.

Rousset F, 2008. Genepop' 007: A complete re-implementation of the genepop software for Windows and Linux. Mol. Ecol. Resour. 8: 103–106.

Smith EJ, Partridge JC, Parsons KN, White EM, Cuthill IC et al., 2002. Ultraviolet vision and mate choice in the guppy *Poecilia reticulata*. Behav. Ecol. 13: 11.

Sobel JM, Chen GF, Watt LR, Schemske DW, 2010. The biology of speciation. Evolution. 64: 295–315.

SPECIATION-Network, 2011. What do we need to know about speciation? Trends Ecol. Evol. 27: 27–39.

Stuart-Fox D, Owens IPF, 2003. Species richness in agamid lizards: Chance, body size, sexual selection or ecology? J. Evol. Biol. 16: 659–669.

Van Doorn GS, Edelaar P, Weissing FJ, 2009. On the origin of species by natural and sexual selection. Science. 326: 1704–1707.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P, 2004. MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. Mol. Ecol. Notes 4: 535–538.

Webb SA, Graves JA, Macias-Garcia C, Magurran AE, Foighil DO et al., 2004. Molecular phylogeny of the livebearing Goodeidae (Cyprinodontiformes). Mol. Phylogenet. Evol. 30: 527–544.

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

Wilson AB, Noack-Kunnmann K, Meyer A, 2000. Incipient speciation in sympatric Nicaraguan crater lake cichlid fishes: Sexual selection versus ecological diversification. P. Roy. Soc. Lond. B. Bio. 267: 2133–2141.

Wolf JBW, Lindell J, Backström N, 2010. Speciation genetics: Current status and evolving approaches. Philos. T. Roy. Soc. B365: 1717.