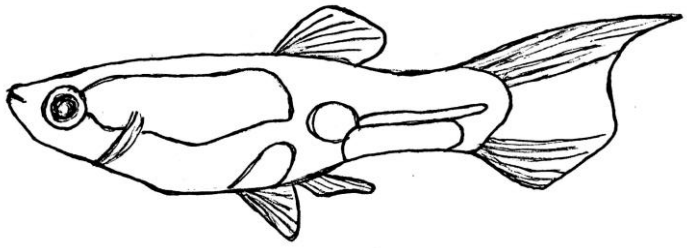
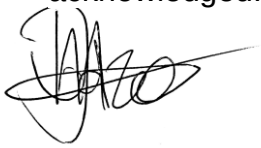
Genomics of adaptation in experimental populations of Trinidadian guppies *(Poecilia reticulata)*

Submitted by **Mijke Janne van der Zee** to the University of Exeter
as a thesis for the degree of
Doctor of Philosophy in Biological Sciences
In September 2020





Signature: ……..

ABSTRACT

Natural populations are increasingly affected by human-mediated changes in the environment. Despite substantial evidence for rapid phenotypic evolution in response to changing environments, the role of genetics underlying rapid adaptation remains relatively unexplored. Natural populations of guppies (*Poecilia reticulata*) in Trinidad show clear repeatable phenotypic adaptation to low- (LP) and high (HP) -predation environments. Experiments where guppies were transplanted from HP to LP environments showed the LP phenotypes can evolve in as little as four years. Using whole genome sequencing of two well-established introductions (sampled at 64 and 114 generations post introduction), and four newly introduced populations (sampled 8-10 generations post introduction), I investigate the genetic response to novel environments. I uncover varying demographic histories between the two established populations, with evidence of bottlenecks in one and extensive population growth in the other. Both experimental populations showed signatures of convergent evolution with a natural LP population, but I found little evidence of convergence between them, indicating they explored different molecular pathways to achieve similar phenotypes. In the second half of the thesis, I show four recently introduced populations evolved minor genetic changes with respect to their HP source, and find no evidence of bottlenecks in three of the four populations. I find signals of selection in all four populations, and a 2Mb region on chromosome 15 showed a consistent signal of selection in three of the four populations. Finally, using a multivariate analysis of allele frequency changes I uncovered subtle parallel changes at multiple loci across all four populations, indicating the approach has potential to detect convergent evolution in rapidly evolving populations, as well as identifying signatures of polygenic selection. Taken together, the findings in

this thesis contribute to a better understanding of the process of rapid genetic adaptation after a sudden environment shift in natural populations.

TABLE OF CONTENTS

TITLE PAGE **1**

LIST OF FIGURES **7**

LIST OF TABLES **9**

DECLARATION **12**

ACKNOWLEDGEMENTS **13**

LIST OF ABBREVIATIONS **15**

CHAPTER 1 GENERAL INTRODUCTION **17**

1.1 Genetic adaptation and selection *17*
1.1.1 Phenotypic adaptation 17
1.1.2 Detecting selection 18
1.1.3 Polygenic selection 20
1.1.4 Population demography in new environments 21
1.1.5 Studying rapid convergent evolution 23

1.2 The Trinidadian guppy as a model system *24*
1.2.1 Biology and ecology of the guppy 24
1.2.2 Northern Trinidad as a model system for convergent evolution 25

1.3 Aims and objectives of this thesis *29*

CHAPTER 2 - NEUTRAL EVOLUTION OF ESTABLISHED INTRODUCTION POPULATIONS OF GUPPIES **33**

2.1 Abstract *33*

2.2 Introduction *33*

2.3 Methods *39*

2.4 Results *48*

2.5 Discussion *63*

CHAPTER 3 – EFFECTS OF STARTING GENETIC VARIATION ON CONVERGENT EVOLUTION IN ESTABLISHED EXPERIMENTAL POPULATIONS OF GUPPIES **70**

3.1 Abstract *70*

3.2 Introduction *70*

3.3 Methods *74*

3.4 Results *77*

3.5 Discussion *108*

CHAPTER 4 – INFERRING DEMOGRAPHIC HISTORIES OF RAPIDLY ADAPTING POPULATIONS **115**

4.1 Abstract *115*

4.2 Introduction *115*
4.3 Methods *119*
4.4 Results *126*
4.5 Discussion *138*

CHAPTER 5 – DETECTING PARALLEL SELECTION IN RAPIDLY EVOLVING EXPERIMENTAL POPULATIONS OF GUPPIES **144**
5.1 Abstract *144*
5.2 Introduction *145*
5.3 Methods *148*
5.4 Results *153*
5.5 Discussion *172*

CHAPTER 6 - GENERAL DISCUSSION **178**
6.1 The importance of demography *180*
6.2 Molecular convergence *182*
6.3 The impact of time on measuring genomic adaptation *185*
6.4 Confounding environmental variables *189*
6.5 Genomic context and the search for signals of selection *192*
6.6 Other avenues for future research *194*
6.7 Concluding remarks *196*

REFERENCES **197**

SUPPLEMENTARY MATERIAL **214**

LIST OF FIGURES

Figure 1.1 Colourful male guppy and two female guppies. 25
Figure 1.2 Maps showing the location of Trinidad and the Northern Range Mountains 26
Figure 2.1 Map showing the locations of the sampled populations 38
Figure 2.2 Schematics of the demographic models tested 48
Figure 2.3 Principal component analysis of all six populations 49
Figure 2.4 Frequency distributions of summary statistics 51
Figure 2.5 Plots showing ROH per size category. 54
Figure 2.6 F_{ROH} per individual per population. 55
Figure 2.7 Likelihood distributions of each model for each HP-LP pair 58
Figure 2.8 Schematic showing migration 59
Figure 2.9 Parameter estimates of current effective populations size 60
Figure 3.1 Distributions of genome scan measures 79
Figure 3.2 Number of overlapping outlier windows among all comparisons in TULP/GHP. 81
Figure 3.3 Number of overlapping outlier windows among all comparisons in ECLP/ECHP. 82
Figure 3.4 Number of overlapping outlier windows among all comparisons in GLP/GHP. 82
Figure 3.5 Distribution of XP-EHH for overlapping outlier positions 83
Figure 3.6 Overlapping outlier windows on chromosome 1 86
Figure 3.7 Per SNP values for the three consecutive outlier windows on chromosome 1 87
Figure 3.8 Overlapping outlier windows on chromosome 20/scaffold 94 88
Figure 3.9 Overlapping outlier windows on chromosome 8 90
Figure 3.10 Per SNP values for the three consecutive outlier windows on chromosome 8 91
Figure 3.11 Proportion of heterozygotes for a region on chromosome 8 92
Figure 3.12 Overlapping outlier windows on chromosome 12 94
Figure 3.13 Per SNP values for the four consecutive outlier windows on chromosome 12 95
Figure 3.14 Proportion of heterozygotes for a region on chromosome 12 96

Figure 3.15 Overlapping outlier windows among outlier windows of GLP/GHP D_{XY} and from all measures of the introduced populations 99
Figure 3.16 Per SNP values for the overlapping outlier window on chromosome 15 101
Figure 3.17 Tajima's D for GHP and the two LP populations in the identified outlier windows. 107
Figure 4.1 Map showing the locations of the introduction experiments. 119
Figure 4.2 Schematics of the demographic models tested with FASTSIMCOAL2. 126
Figure 4.3 Principal component analysis of all six populations 128
Figure 4.4 Frequency distributions of summary statistics 130
Figure 4.5 Plots showing ROH per size category. 133
Figure 4.6 F_{ROH} per individual per population. 134
Figure 4.7 Likelihood distributions of each model for each HP-LP pair 136
Figure 5.1 Schematic showing parallel, non-parallel and anti-parallel changes 151
Figure 5.2 Distributions of minor allele frequency changes between GHP and each of the LP populations. 154
Figure 5.3 Genome-wide values of haplotype genome scans. 156
Figure 5.4 Overlap of genome scan outliers among the four introduced populations 157
Figure 5.5 Allele frequency changes in XP-EHH outlier windows 161
Figure 5.6 Allele frequency changes in iHH12 outlier windows 162
Figure 5.7 Eigenvalues for the first two eigenvectors along the genome 163
Figure 5.8 Eigenvalues for the first eigenvector along chromosomes with the most outlying windows. 164
Figure 5.9 Allele frequency changes in outlier windows on eigenvector 1. 165
Figure 5.10 Per window values of XP-EHH, iHH12 and $|\Delta AF|$ on chromosome 15 168
Figure 5.11 Per SNP values of XP-EHH, iHH12 and $|\Delta AF|$ 169
Figure 5.12 Linkage disequilibrium on chromosome 15 171

SUPPLEMENTARY FIGURES

Figure S3.1 LD decay analysis to determine window size for the genome scan analyses. Vertical dashed line indicates the selected window size, horizontal dashed line the cutoff for LD decay……224

LIST OF TABLES

Table 2.1 Summary of population information..... 41
Table 2.2 Pairwise global F_{ST} values for each population. 50
Table 2.3 Summary of population genetic statistics. 52
Table 2.4 The average length and number of ROH 54
Table 2.5 Δ-likelihoods for the best runs of each model in FASTSIMCOAL2. 57
Table 2.6 FASTSIMCOAL2 parameter estimates..... 61
Table 3.1 Summary of genome scan results for ΔAF, F_{ST}, $\Delta\pi$ and D_{XY}..... 80
Table 3.2 Observed heterozygosity in the outlier regions 89
Table 3.3 Number of observed outliers compared among populations. 97
Table 3.4 Distribution of overlapping outlier windows across the genome for ΔAF among populations. 98
Table 3.5 Positions of overlapping outlier windows among three of the four measures..... 100
Table 3.6 Tajima's D in the overlapping outlier windows 102
Table 4.1 Summary of population information..... 121
Table 4.2 Pairwise global F_{ST} values among all six populations. 129
Table 4.3 Summary of population genetic statistics. 131
Table 4.4 The average length and number of ROH 132
Table 4.5 Δ-likelihoods for the best runs of each model in FASTSIMCOAL2. 137
Table 5.1 Summary of allele frequency changes in all populations..... 154
Table 5.2 Overlapping outlier windows among all populations for XP-EHH and iHH12. 159
Table 5.3 Overlapping outlier windows collapsed into regions of at least three consecutive windows for XP-EHH. 160
Table 5.4 Overlapping outlier windows collapsed into regions of at least three consecutive windows for iHH12..... 160

Table 5.5 Overlapping outlier windows between eigenvector 1 outliers and iHH12 outlier windows..... 166
Table 5.6 Overlapping outlier windows between eigenvector 1 and XP-EHH outlier windows. 167
Table 6.1 Overlapping outlier regions among TULP and the new introduction populations and among ECLP and the new introduction populations..... 184
Table 6.2 Ecological treatments per population in newly introduced populations 192

SUPPLEMENTARY TABLES

Table S2.1 Fastsimcoal2 GHP/GLP parameter ranges used in the simulations of model 1 with migration……..……212
Table S2.2 Fastsimcoal2 ECHP/ECLP parameter ranges used in the simulations of model 5 with migration……...212
Table S2.3 Fastsimcoal2 GHP/TULP parameter ranges used in the simulations of model 1 without migration……213
Table S2.4 Pairwise global F_{ST} values among all six populations with outlier windows removed from the data set. Mean F_{ST} below the diagonal, median F_{ST} above the diagonal. Population names can be found in table 2.1…….213
Table S2.5 Pairwise global F_{ST} values among all six populations with the sex chromosome removed from the data set. Mean F_{ST} below the diagonal, median F_{ST} above the diagonal. Population names can be found in table 2.1……..213

Table S3.1 Positions of outlier windows overlapping among three of the four measures in GHP/TULP……200
Table S3.2 Positions of outlier windows overlapping among three of the four measures in ECHP/ECLP…….201
Table S3.3 Positions of outlier windows overlapping among F_{ST}, $|\Delta AF|$ and D_{XY} measures in GHP/GLP……205

Table S4.1 Fastsimcoal2 parameter ranges used in the simulations for all four populations……..224
Table S4.2 Pairwise global F_{ST} values among all six populations with outlier windows removed from the data set……224
Table S4.3 Pairwise global F_{ST} values among all six populations with the sex chromosome removed from the data set……224
Table S4.4 FASTSIMCOAL2 parameter estimates from the best run and 95% confidence intervals generated by simulations of the best model in ILL……..225

Table S5.1 XP-EHH outlier windows for ILL……226
Table S5.2 XP-EHH outlier windows for IUL……227
Table S5.3 XP-EHH outlier windows for IC……...227
Table S5.4 XP-EHH outlier windows for IT……228
Table S5.5 iHH12 outlier windows for ILL……229
Table S5.6 iHH12 outlier windows for IUL……229
Table S5.7 iHH12 outlier windows for IC.....230
Table S5.8 iHH12 outlier windows for IT.....231
Table S5.9 Windows with outlier eigenvalues on eigenvector 1.....232
Table S5.10 Windows with outlier eigenvalues on eigenvector 2……..233

DECLARATION

The work presented in this thesis is my own, however certain people made contributions within various chapters:

Sampling of all fish was done by my supervisor, Bonnie Fraser, in 2011 and 2013.

DNA extraction and library preparation for six of the 10 populations was done by Bonnie Fraser, prior to my arrival. Josephine Paris assisted with the DNA extraction and library preparation of the remaining 4 populations.

Josephine Paris collaborated on the demographic analysis with FASTSIMCOAL2 (chapters 2 and 4). James Whiting assisted with the development of the multivariate analysis of allele frequencies (chapter 5).

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Bonnie for giving me the opportunity to work on this incredible project. You have been an amazing supervisor, giving me space to figure things out on my own, but also be there to provide advice and guidance when I needed it. Your supervision style has kept me motivated and happy throughout the whole project, and I still like the gups after four years!

A thank you to my second supervisors, Adam at Sussex and Jamie at Exeter, for sharing their experience of how to successfully plan and execute a PhD.

I am grateful to the School of Biology, University of Sussex, and to the College of Life and Environmental Sciences, University of Exeter for funding my PhD.

Thank you to the people behind the guppy project, who not only thought up the experiment and provided samples, but also offered valuable insights from their extraordinary knowledge of the system, David Reznick, Joe Travis, Tim Coulson and Ron Bassar.

A big thank you to all members of the Fraser Lab, Bonnie, Josie, Jim, Jenn and Paul. Thanks for all your help with analyses and lab work. Thanks for the heated discussions (cats or dogs? McCartney or Lennon?) and pub trips over the years. I'm grateful for being part of such an amazing group of people!

Josie, my PhD experience would not have been the same without you! Thank you for always being there to help. I will never forget warming up extraction kits in our lab coats, and our days in the fastsimcoal bunker. Thank you for your friendship outside of work. Our countless chats over coffee (or wine), dinners with Elliot and Joan and dance parties (around Xini) in your living room have gotten me through it all!

Jim, thank you for your endless patience answering all my questions. Thank you for your maths-y brain to help me think through some tough analyses and

showing me the right way to use R. Your sense of humour and bottomless knowledge on so many random topics have left me laughing every time. Also, thinking shears will forever be a thing now.

A thank you for the EBE group at Sussex for making the first two years of my PhD so much fun. Friday 5 o'clock songs in the office followed by cheesy chips and beers in The Swan are some of my favourite Brighton moments! And thanks to my fellow Cave Dwellers at Exeter for making me feel at home in our beloved Cave (I wonder what it looks like now!).

Lieve Anne, Devi en Nina! Bedankt voor jullie bezoekjes hier. Het voelde altijd alsof ik ook een beetje op vakantie was als jullie er waren! Bedankt voor de gezellige avondjes wijn in Nederland, hopelijk houden we de traditie om samen mijn verjaardag te vieren er nog vele jaren in!

Pap, mam en broers, ontzettend bedankt voor alle fijne en gezellige momenten in Nederland! Mam, onze dagen samen in Engeland zijn altijd zo fijn. Wandelen, puzzelen, wijntjes drinken en charity shoppen zorgden er altijd voor dat ik even echt kon relaxen. Pap, bedankt voor de biologie inspiratie van jongs af aan. En bedankt voor je hulp met het lezen van hoofdstukken, er zijn niet veel PhD'ers die kunnen zeggen dat hun vader daarmee kan helpen!

Last but definitely not least. Liefste Elliot, thank you so much for everything you have done for me the last three years. Letting me talk genetics at you has helped me solve problems so many times, you are my favourite rubber duck. Thank you for always managing to calm me down when I was feeling overwhelmed. Thank you for taking me on walks and camping trips to get me my dose of nature therapy, I can't wait to go on countless more adventures in our new van!

LIST OF ABBREVIATIONS

ABC	Approximate Bayesian computation
AF	Allele frequency
bp	base pairs
BN	Bottleneck
CI	Confidence interval
DNA	Deoxyribonucleic acid
DNM	*De novo* mutation
ECHP	El Cedro high predation
ECLP	El Cedro low predation
F_{ROH}	Inbreeding coefficient
F_{ST}	Fixation index
GHP	Guanapo high predation
GLP	Guanapo low predation
H_e	Expected heterozygosity
HP	High predation
IBD	Identity-by-descent
IBS	Identity-by-state
IC	Caigual introduced
ILL	Lower Lalaja introduced
IT	Taylor introduced
IUL	Upper Lalaja introduced
LD	Linkage disequilibrium
LP	Low predation

MAD	Median absolute deviation
Mb	Mega bases
N_E	Effective population size
PCA	Principal component analysis
RAD-seq	Restriction site associated DNA sequencing
ROH	Runs of homozygosity
SFS	Site frequency spectrum
SGV	Standing genetic variation
SNP	Single nucleotide polymorphism
TD	Tajima's D
TUHP	Turure high predation
TULP	Turure low predation
VCF	Variant Call Format
WGS	Whole genome sequencing

Chapter 1 GENERAL INTRODUCTION

1.1 Genetic adaptation and selection

Natural ecosystems are increasingly affected by human activities (Rockström *et al.*, 2009). Changes in land use, pollution, hunting and human-induced climate change are already having observable effects on biodiversity, including the composition of biological communities, distribution and phenotypes of species (Catullo *et al.*, 2019). The speed at which these changes occur are unprecedented, and understanding how natural populations respond to the direct and indirect consequences of human-mediated shifts in the environment have become an increasingly important research challenge in contemporary evolutionary ecology (Lavergne *et al.*, 2010).

1.1.1 Phenotypic adaptation

To cope with sudden shifts in the environment, and to avoid local extinctions, populations have to either expand their distributions in search of better suited habitat or become adapted to their new local conditions (Anderson *et al.*, 2012). Phenotypic adaptation is the result of the interactions between several factors: plasticity (Baldwin, 1896), gene flow (Garant *et al.*, 2007) and natural selection (Schluter, 2000). When encountering a new environment, at first adaptation may arise through phenotypic plasticity if the new conditions are within the species/population tolerances (Bradshaw, 2006). It is possible for the newly adapted phenotype to then become heritable via natural selection, a process referred to as genetic assimilation (Waddington, 1953; Price *et al.*, 2003).

Alternatively, selection can act on *de novo* mutations or standing genetic variation (SGV) resulting in genomic adaptation. Adaptation through *de novo*

mutations requires time for beneficial alleles to arise, therefore rapid adaptation is more likely to occur through selection on SGV as the beneficial allele is already present in the population (Barrett and Schluter, 2008). Furthermore, standing variation may have already been tested by selection in past environments, increasing the chance the allele is advantageous (Barrett and Schluter, 2008). Finally, ongoing gene flow and population admixture can also introduce locally adapted alleles into a population. This sharing of adaptive alleles is thought to be especially important for natural populations with reduced SGV, as it allows for rapid adaptation without having to wait for *de novo* beneficial mutations to arise (Hamilton and Miller, 2016).

Under the right conditions, adaptation can happen very rapidly, and recent studies have highlighted cases where adaptive phenotypic change occurred within just a few generations. Examples include beak morphology in Darwin's finches (Grant and Grant, 2008), limb morphology of *Anolis* lizards (Losos *et al.*, 1997), soapberry bug beak length (Carroll and Boyd, 1991), and life histories of guppies (Reznick *et al.*, 1990; Reznick, Bassar, *et al.*, 2019). However, the genomic basis of rapid adaptation is still relatively unknown. With the advent of whole genome sequencing, it has now become possible to identify individual loci under recent selection. By identifying specific loci, it would be possible to link the genotype and environment, and predict the resilience of a genotype or population to a certain environmental change. This could assist in improving management decisions for conservation purposes (Li *et al.*, 2014)

1.1.2 Detecting selection

As natural selection acts on phenotypic traits, the population shifts toward a new optimum by changes in allele frequencies underlying the selected traits

(Hoban *et al.*, 2016). When a beneficial allele increases in frequency in a population because of natural selection, it also affects variation at neighbouring sites as a result of the “hitchhiking effect” (Smith and Haigh, 1974). Such selective sweeps reduce the amount of genetic variation, increase the amount of linkage disequilibrium (LD), and cause an excess of rare alleles and high-frequency-derived alleles (Braverman *et al.*, 1995; Fay and Wu, 2000; Kim and Nielsen, 2004). To detect these signatures of selection, three measures are commonly used: 1) linkage disequilibrium (LD), 2) the site frequency spectrum (SFS) and 3) population differentiation-based tests. Most LD-based methods identify long homozygous regions with high frequencies of certain haplotypes (Sabeti *et al.*, 2002; Garud *et al.*, 2015), or search for specific spatial patterns of LD caused by independent recombination events (Kim and Nielsen, 2004). Selective sweeps leave a signature of increased low- and high-frequency variants that can be detected either by tests based on the population mutation rate (Tajima, 1989a; Fu and Li, 1993), or comparing the SFS of candidate genomic regions with neutral assumptions on SFS or SFS based on the whole data set (Nielsen, 2005). Finally, population differentiation tests assume that populations in different environments experience different selective regimes. Beneficial alleles are expected to be present in high frequencies in the population experiencing positive selection, whereas they should occur at low or intermediate frequencies in the other population (Weigand and Leese, 2018).

Historically, the focus of adaptation studies has been on selection on one or a few genetic loci, where a new mutation rapidly increases in frequency and becomes fixed in the population. This type of selective sweep leaves the classic selective signature described above, and most available methods aim to detect these “hard sweeps”. Recently however, studies have suggested that instead

adaptation through “soft” sweeps might be more common (Pritchard *et al.*, 2010). Soft sweeps can be caused by 1) a beneficial allele entering the population multiple times via migration of different individuals, 2) different mutations in the same genomic region that have a similar selective advantage or 3) selection acting on SGV when a change in the environmental conditions leads to a selective advantage of a previously slightly deleterious or neutral genetic variant (Weigand and Leese, 2018). As a result of any of these causes, the selected variant in a soft sweep is not associated with a single genomic background, but instead there are multiple copies of the beneficial allele and therefore multiple haplotypes increase in frequency during the selective sweep (Hermisson and Pennings, 2005). This results in levels of genetic diversity and allele frequency spectra that are less perturbed than in a hard selective sweep (Barrett and Schluter, 2008; Garud *et al.*, 2015), and therefore difficult to detect with many of the standard methods.

In rapidly evolving populations, adaptation is predicted to occur through soft selective sweeps, as the beneficial alleles are immediately available for selection to act upon after a change in the environment (Hermisson and Pennings, 2005; Barrett and Schluter, 2008), compared to *de novo* mutations that take time to arise. Detecting the adaptive loci in the early stages of adaptation is an important aspect of evolutionary studies, as it can help understanding which regions of genome are the first to respond in a changing environment.

1.1.3 Polygenic selection

Most population genetic models for detecting selection only consider one or a few loci at which positive selection acts (Orr, 2005; Stephan, 2016). This is in contrast with classical models in quantitative genetics, where it is assumed that most traits

of interest are highly polygenic and adapt through subtle allele frequency shifts at many loci at a time (Pritchard and Di Rienzo, 2010). For many traits, selection on SGV at many loci simultaneously would allow for more rapid adaptation than selection at one or a few loci (Pritchard *et al.*, 2010). If adaptation occurs through polygenic adaptation, many selected alleles will change in frequency simultaneously, but generally not to fixation, and this makes detecting polygenic selection very difficult with classic population genetic methods (Chevin and Hospital, 2008; Coop *et al.*, 2009). Jain and Stephan (2017) suggest that analysing allele frequency shifts at all loci simultaneously (instead of examining individual SNPs) might provide a useful approach to investigate signals of polygenic selection. Currently however, methods to detect polygenic selection using WGS data are still rare.

1.1.4 Population demography in new environments

Changes in population size and other demographic events can result in patterns of diversity that resemble those observed in a selective sweep. Rapid population expansion for example, results in an increase of rare alleles that is also observed during a selective sweep (Fu and Li, 1993). Depending on the strength of the bottleneck, different changes in diversity might be expected. A strong bottleneck resembles a selective sweep with fixation before any recombination has occurred between the selected and neutral loci, which would remove all variation around the selected locus (Depaulis *et al.*, 2003). The pattern resulting from a more moderate bottleneck on the other hand, could be confused with the pattern left behind by genetic hitchhiking, where recombination between the selected and neutral loci occurred, or it may reflect an incomplete selective sweep. Both scenarios (strong and moderate bottlenecks) show that it is difficult

to distinguish selective sweeps from random demographic events (Depaulis *et al.*, 2003). To address this issue, we investigate signals of selective sweeps in multiple populations, allowing us to distinguish drift from selection.

Population bottlenecks reducing genetic variation are predicted to occur in populations that have recently been introduced to a new environment (Nei *et al.*, 1975), however there is growing evidence that many species do not show evidence of genetic bottlenecks (i.e. reduced genetic diversity, small population sizes) after settling in a new environment (Dlugosch and Parker, 2008). We currently know little about why different populations respond differently to an introduction event. By comparing distinct introductions with varying demographic histories, my thesis aims to investigate the effects of demography on evolvability after an introduction event, and assess how these histories affect the occurrence of genomic convergence.

Besides changes in population size, migration and gene flow also affect our ability to detect selection. Generally, the assumption is that in scenarios without migration or very low levels of migration, a beneficial allele becomes fixed in populations experiencing positive selection, whereas in the alternative environment the allele remains at low/intermediate frequencies. However, in the presence of moderate to high levels of migration, the beneficial allele under selection in one environment can end up swamping the population in the environment not experiencing selection, reducing the absolute values of differentiation measures to levels that are not necessarily different from neutrality (Whiting and Fraser, 2020).

1.1.5 Studying rapid convergent evolution

Populations that have independently evolved convergent phenotypes in response to similar changes in the environment provide an excellent opportunity to study the genomic basis of rapid adaptation. Processes other than selection are unlikely to result in repeated evolutionary changes among independent populations, therefore instances of convergent evolution provide strong evidence for natural selection. This aspect makes convergent evolution such an intensely studied subject, as it creates an opportunity to distinguish selective events from random genetic drift, and study the mechanisms of genomic adaptation. In this thesis, I adopt the framework of Arendt and Reznick (2008) and refer to all cases of repeated evolution as convergent.

Investigating the genomic basis of convergent evolution can be done in several ways. First, using comparative studies of populations that have experienced similar environments. For example, the evolution of pitcher plants in several unrelated genera in response to a nutrient-poor environment (Thorogood *et al.*, 2018), the loss of armour plating in multiple populations of marine stickleback adapting to freshwater (Jones *et al.*, 2012), or the evolution of light coat colour in beach mice (Steiner *et al.*, 2009). However, many known examples of convergent evolution have been evolving for thousands of years, and we do not know whether convergence occurred in the earliest generations of adaptations or over longer timescales, limiting our understanding of the fundamental genomic architecture underlying convergent phenotypes.

A second approach to study rapid convergent evolution is provided by the experimental evolution framework. Experimental evolution gives the researcher control over the environment and it can be set up to include multiple replicate populations experiencing the same environmental change (Kawecki *et al.*, 2012).

Combined with whole genome sequencing (WGS) it allows us to track allele changes in real-time and investigate the genomic basis of rapid adaptation (Stern, 2013). Furthermore, experimental evolution is not dependent on finding systems with recent independently evolved convergent phenotypes. Even though experimental studies have greatly increased our understanding of convergent evolution, they are often limited to model species and laboratory strains often with low amounts of genetic variation (Fraser and Whiting, 2019). These types of populations will have markedly different demographic parameters and genomic context from natural populations, and thus the dynamics of adaptation in these populations might not be representative of those in natural populations (Wood *et al.*, 2005). Experimental evolution studies combined with WGS data using wild populations are still rare, but provide an excellent opportunity to investigate the effects of complex natural processes on early adaptation to a new environment. Examples of experimental natural populations can be found in the guppy system in northern Trinidad.

1.2 The Trinidadian guppy as a model system

1.2.1 Biology and ecology of the guppy

The guppy, *Poecilia reticulata*, is a small tropical freshwater fish displaying sexual dimorphism. Males are smaller than females and have conspicuous colour patterns on the body (figure 1.1). Guppies are native to northern South America and parts of Central America; naturally occurring in Trinidad, Venezuela, Guyana and Surinam (Magurran, 2005). Guppies can be found on every continent however, except Antarctica, as a consequence of the pet trade and because they were introduced as a means of controlling mosquitoes (El-Sabaawi *et al.*, 2016). Guppies are placed in the Poeciliidae family, and the subfamily Poeciliinae.

Figure 1.1 Colourful male guppy and two female guppies. (https://commons.wikimedia.org/wiki/File:Guppy_pho_0048.jpg).

The reproductive cycle of the guppy lasts 25-30 days, and continues throughout the year with some seasonal variation (Haskins *et al.*, 1961; Reznick, 1989; Alkins-Koo, 2000). Like all members of the poeciliid family, guppies have internal fertilisation (Wourms, 1981), and sperm transmission to females is done using the gonopodium, a modified anal fin that serves as a copulatory organ. Female guppies can store sperm in the folds of their ovaries and gonoducts for up to eight months (Winge, 1937), and usually mate multiple times, with a median of two sires per brood (Becher and Magurran, 2004). Guppies are lecithotrophic livebearers, meaning embryos are supported by yolk that females deposited prior to fertilisation (Constanz, 1989). Female guppies first produce offspring around 10-20 weeks, with 2-3 generations per year, while male maturation can be reached in 7 weeks.

1.2.2 Northern Trinidad as a model system for convergent evolution

The Northern Range Mountains in Trinidad have long been one of the prime models for studying convergent evolution in vertebrates. In 1961, Caryl Haskin's commented that the rivers of the Northern Range resemble a natural laboratory because of the parallel rivers draining the slopes (Haskins *et al.*, 1961). The rivers

drain into three separate drainages: the Caroni drainage flows west into the Gulf of Paria, the Oropuche drainage flows east to the North Atlantic Ocean, and the northern drainage drains a set of parallel rivers from the northern flanks of the mountain into the Caribbean Sea (figure 1.2). Many of these rivers are punctuated by waterfalls forming natural barriers that prevent larger fish species to migrate upstream, including potential predators of the guppy. Stretches of river downstream of these waterfalls have rich fish assemblages that include guppies, *Rivulus hartii* (Hart's rivulus), *Crenicichla alta* (pike cichlid), *Hoplias malabaricus* (wolf fish), *Aequiedens pulcher* (blue acara), *Astynax bimaculatus* (two-spot sardine), *Hemibrycon taeniurus* (mountain sardine), *Synbranchus marmoratus* (swamp eel) and *Awaous taiasica* (sand fish). Above the waterfalls, generally only guppies, *R. hartii* and, occasionally *A. pulcher* are found. Of the species mentioned, Haskins *et al*. (1961) determined *C. alta, H. malabaricus, A. bimaculatus* and *A. pulcher* to be the more serious guppy predators, resulting in a contrast in predation risk between the downstream high predation (HP) sites and the upstream low predation (LP) sites. The predation gradient between upriver and lowland streams is replicated across independent rivers.

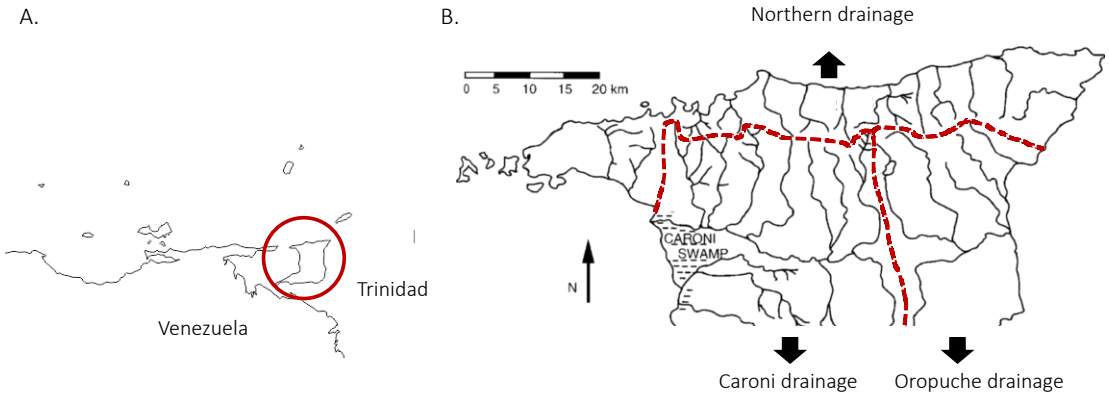


Figure 1.2 Maps showing the location of Trinidad and the Northern Range Mountains A) Location of Trinidad on northern coast of South America. B) The three drainages of the Northern Range Mountains.

Some aspects of the rivers, such as flow regime, water depth and substratum, are usually similar directly above and below the waterfalls/barriers (Reznick *et al.*, 2001). However, there are some additional ecological factors that consistently differ between upstream and downstream localities that could confound the HP-LP dichotomy. LP sites tend to have denser canopy cover compared to HP localities, leading to a decrease in light reaching the river surface which results in a reduction of primary productivity in the upper stretches of the rivers (Reznick *et al*., 2001). Grether *et al*. (2001) found that a reduced canopy cover in HP sites led to increased algal standing crops and a different composition of algal species that may have different nutritional values compared to those found in LP environments. Reznick *et al*. (2001) also found that HP environments had fewer large, old fish and more young, small fish than LP environments, resulting in a lower biomass per unit area in HP sites, which should result in less competition for food. Furthermore, HP populations have lower guppy densities than LP locations (Reznick and Endler, 1982).

Researchers have documented striking convergent phenotypes among the LP populations in different rivers compared to their HP counterparts. Reznick and Endler (1982) found that guppies from LP sites across the system matured at a large size, reproduced less frequently and produced fewer but larger offspring compared to guppies from HP sites. Reznick (1982) confirmed a genetic basis for these traits when he found life history patterns persisted after two generations in a common environment. Furthermore, males from LP sites tend to be more colourful than their HP counterparts, with more spots and larger individual spots (Endler, 1980). Finally, there are also behavioural differences between HP and

LP guppies. Males and females in LP sites are less wary of predators (Kelley and Magurran, 2003), and shoal less (Seghers, 1974) than HP fish.

To investigate the role of selection for these life history patterns, researchers conducted some of the first translocation experiments with natural populations. In total, there have been four experiments where guppies from HP localities were introduced to previously guppy-free LP locations across the Northern Range of Trinidad. Three of these experiments have been monitored for over thirty years; they were introduced in 1956, 1976, and 1981 (Haskins *et al.*, 1961; Endler, 1980; Reznick and Bryga, 1987). The fourth experiment was conducted in 2008 and 2009, and involves four replicate populations originating from the same HP source in the Guanapo river. Monitoring these experimental populations revealed they evolved a phenotype similar to those found in naturally colonised LP populations. Male guppies in LP sites matured at a later time and larger size, and they had increased ornamental colouration compared to HP source population. All these studies found that these adaptations can be quite rapid, with some responses occurring within four years of the introduction (Endler, 1980; Reznick and Bryga, 1987; Kemp *et al.*, 2018; Reznick, Bassar, *et al.*, 2019).

Recently, the first genome scan of guppy populations has been conducted to examine the molecular basis of convergent phenotypic evolution in the guppy (Fraser *et al.*, 2015). They found limited molecular convergence of selected regions in natural LP populations, whereas the introduced populations had several regions across the genome with signatures of selection common to all populations (Fraser *et al.*, 2015). This was attributed to differences in starting genetic variation, as natural LP populations are likely colonised by few individuals, whereas the introduced populations are founded by large numbers of guppies. This study also found that natural populations experienced population

growth since colonisation whereas the introduced populations shrank. However, this study was based on just a few populations, and used restriction site associated DNA sequencing (RAD-seq) data, which has limited power to detect loci under selection due to the low density of markers (Lowry *et al.*, 2017). By examining more populations and using higher resolution genome data, it would be possible to get a clearer picture of the genomic basis of rapid adaptation in the guppy system. Also, by analysing the demographic histories of more introduced populations we can distinguish the effects of demographic history on molecular convergence.

In summary, the Trinidadian guppy system provides an excellent opportunity to study evolution in natural populations. The short generation time of guppies allows researchers to answer questions about the evolutionary process in a relatively short amount of time. The replicated nature of HP/LP environments combined with the experimentally introduced populations provide a unique opportunity to investigate the process of rapid adaptation to a changing environment.

1.3 Aims and objectives of this thesis

The studies in this thesis are aimed at increasing our understanding of the genetic basis of rapid adaptation in Trinidadian guppies. I focus on several experimentally introduced populations of guppies to explore patterns of diversity and demographic histories before investigating signals of selection and the occurrence of molecular convergence among populations using a whole genome approach. Importantly, through analysis of the data chapters presented here, I address subjects such as:

The importance of demography

Did the introduction experiments cause genetic bottlenecks in the populations? Did the initial genetic variation play a role in the response after translocation? Can we infer demographic histories of very young (<20 generations) populations? How does demography affect selection? And does demography affect the occurrence of molecular convergence?

Molecular convergence

Is the phenotypic convergence we observe in guppy populations the result of molecular convergence? How do differences in available genetic variation affect the occurrence of molecular convergence? Do we observe higher levels of molecular convergence in recently established populations compared to established introduction populations, because selection acts on SGV? Do experimental populations experience more molecular convergence than natural populations?

Rapid adaptation

What is the genomic signature of rapid adaptation? How does time affect our ability to detect selection? Can available methods detect selection in very recent populations?

Chapter 2: Neutral evolution of established introduction populations of guppies

I use genome-wide sequencing data to investigate neutral genomic variation in two long-term introduction populations, their sources, and a natural HP-LP pair in the same drainage. Demographic processes can leave traces in the genome

resembling those left behind by selective events, and they can influence the course of natural selection. Therefore, it is important to understand the demographic history of a population before investigating patterns of selection. In this chapter, I infer demographic histories of the six populations using patterns of neutral genetic variation, runs of homozygosity, and by coalescent modelling of each population, in order to investigate how the introductions have affected neutral variation in the two experimental populations.

Chapter 3: Effects of starting genetic variation on convergent evolution in established experimental populations of guppies

In this chapter, I use information about differences in the amounts of genetic variation obtained in Chapter 2 to investigate how these differences affect selection within a river, and how they affect the likelihood of convergent molecular evolution in two established populations of guppies. I use an outlier window approach with multiple measures to identify putatively selected loci between HP/LP pairs.

Chapter 4: Inferring demographic histories of rapidly evolving populations of guppies

In this chapter, the focus shifts to investigating patterns and processes involved in very recent adaptation. Using whole genome sequencing (WGS) data from four recently (<20 generations) introduced populations, I analyse patterns of neutral genomic variation to assess if the populations experienced a bottleneck after the introduction. Furthermore, the available information for these populations from previous work (Kemp *et al.*, 2018; Reznick, Bassar, *et al.*, 2019)

is used to investigate whether coalescent modelling can correctly infer demographic histories of recently established populations.

Chapter 5: Detecting parallel selection in rapidly evolving experimental populations of guppies

In the final data chapter, I investigate signals of selection in four rapidly evolving introduction populations. In addition to two genome scans based on haplotype homozygosity, I adapt a newly developed method to detect more subtle changes in allele frequency among all four populations to analyse which genomic regions show parallel changes.

Chapter 2 - NEUTRAL EVOLUTION OF ESTABLISHED INTRODUCTION POPULATIONS OF GUPPIES

2.1 Abstract

Understanding the process of adaptation at the genomic level is one of the main focus points of evolutionary biology. However, in order to accurately interpret patterns of differentiation across the genome it is important to understand a population's demographic history. Demographic processes such as bottlenecks, migration and inbreeding affect genetic variation underlying traits under selection. Here, we explore these processes in populations of Trinidadian guppies, *Poecilia reticulata*. Using whole-genome sequencing data from one naturally adapted high- (HP) and low- (LP) predation pair and two experimentally established populations and their sources, we uncover varying demographic histories among the populations. We find the two experimental populations represent different scenarios. One that is older and likely had high levels of genetic diversity at the time of introduction, resulting in extensive population growth. The other a younger population, originating from a source with reduced genetic variation, which led to further a reduction of genetic variation and limited population growth. Our results show demographic histories can hugely affect the amount of genetic variation available in populations, which in turn influences the efficiency of selection and the probability of convergence at the molecular level.

2.2 Introduction

One of the main goals in the field of population genomics is to investigate signatures of selection across the genome. However, accurate interpretations about how and whether natural selection has occurred in a population depend

heavily on its demography. First, demographic processes resulting in a change in population size, such as bottlenecks, migration, inbreeding, and periods of exponential growth, can leave traces in the genome similar to selection. For example, both selective sweeps and population bottlenecks result in a pattern of reduced genetic diversity (Tajima, 1989a; Galtier *et al.*, 2000). Second, the course of natural selection can differ depending on demography, as it influences the efficacy of selection and determines the amount of variation on which selection can act (Rosenblum *et al*., 2014).

There are multiple ways to infer demographic histories from genomic data, broadly these are divided into two general approaches: 1) methods based on the site frequency spectrum (SFS), which encompasses summary statistics, Bayesian computation (e.g. ABC, Beaumont, Zhang and Balding, 2002), and coalescent simulation (e.g. fastsimcoal2, Excoffier *et al.*, 2013) 2) haplotype patterns (Beichman *et al.*, 2018). The SFS is a summary of genome-wide genetic polymorphism within and between populations, and is influenced by the history of a population. Using coalescent theory (Kingman, 1982), we can infer how different demographic events change the SFS by simulating the predicted SFS for a particular demographic model, and then asses its fit to the observed SFS (Beichman *et al.*, 2018). The SFS assumes all sites are independent and demographic inference methods based on the SFS do therefore not incorporate linkage between sites. Including information about linkage however, can be especially informative when investigating recent demographic histories. Haplotype-based methods analyse the abundance and length of homozygous tracts across the genome, thus considering linkage between sites. These runs of homozygosity (ROH) arise when there is a limited effective population size (N_E) or through consanguineous mating (Ceballos *et al.*, 2019). ROH break down as

a function of time, mutation rate and recombination rate (Brüniche-Olsen *et al*., 2018), and therefore exist in a continuum of homozygous lengths, with shorter segments inherited from more distant common ancestors and longer segments from recent ancestors (Kirin *et al.*, 2010). ROH are found in every individual, so by analysing the continuum of ROH, it is possible to infer individual ancestry.

Methods like these offer powerful ways to investigate the processes shaping patterns of genetic diversity in populations. The guppy system in northern Trinidad provides a unique resource to apply these methods. Several long-term introduction experiments in this system have previously shown that guppies experienced selection after being transplanted from a high predation to a low predation environment (Endler, 1980; Reznick and Bryga, 1987; Kemp *et al.*, 2009). Here, by leveraging the available information about these introduction experiments (time of introduction and number of individuals introduced), I tested the inference of the demographic models. The knowledge we gain from this can subsequently be used to account for demography when searching the genome for signals of selection, the subject of the next chapter.

The Trinidadian guppy system

The guppy system in the Northern Range Mountains of Trinidad is a well-known model for studying phenotypic evolution in wild and experimental populations. The mountains are drained by a set of parallel rivers. Many of these rivers have waterfalls that prevent upstream colonisation of larger fish species, including guppy predators *Crenicichla alta* and *Hoplias malabaricus* (Magurran, 2005). Above the waterfalls there is a relatively low predation (LP) environment, whereas downstream of the barriers is considered a high predation (HP) environment. Males from LP environments are more colourful, have a larger

number of spots and larger individual spots than their HP counterparts (Endler, 1980). Additionally, male and female LP guppies have larger body sizes at maturation and reproduce less frequently than guppies occupying HP environments (Reznick & Endler, 1982). This pattern is broadly replicated in rivers across the Northern Range Mountains, but there is variation in some aspects of phenotypes from similar environments (Endler & Houde, 1995; Kemp *et al*., 2009). Many of the phenotypic differences between HP and LP populations are also heritable in laboratory experiments (Reznick and Endler, 1982; Reznick and Bryga, 1996), suggesting convergent phenotypic evolution has a genetic basis. This makes the guppy system ideal to investigate convergence at the molecular level.

Another advantage of this system are several long-term introduction experiments that were conducted to study the rate of adaptive evolution in guppies. Here, guppy populations transplanted from HP to LP environments evolve phenotypes typical of natural LP populations. Both male and female guppies mature at a later age and are larger when they reach maturity, males are more colourful than their HP ancestor, and females produce fewer and larger offspring (Endler, 1980; Reznick and Bryga, 1987; Kemp *et al.*, 2009). In this chapter, the focus is on two such experiments; one made in 1957 by Haskins in the Turure River (1961), and the second by Reznick and Bryga in the El Cedro River in 1981 (1987).

Guppy introduction experiments in the wild

Turure River

When Shaw *et al*. (1991) investigated allozymic differences among guppy populations in the Northern Range Mountains of Trinidad, they found a population

from Turure river in the Oropuche drainage that was more similar to populations from the Caroni drainage than to other populations in the Oropuche at 25 allozyme loci. Personal communication with Caryl Haskins revealed the introduction of 200 guppies from a HP site in the Arima in the Caroni drainage into a guppy free LP site in an upstream location of the Turure River in the Oropuche drainage in 1957 (TULP, figure 2.1). Later studies revealed that the introduced population had spread downstream, introgressing with native populations of several HP populations in the Turure River (including the HP population considered here, TUHP), and that TULP and TUHP are more closely related to Guanapo HP (GHP) and Guanapo LP (GLP) in the Caroni drainage than to other rivers in the Oropuche drainage (Shaw *et al.*, 1992; Becher and Magurran, 2000; Suk and Neff, 2009; Willing *et al.*, 2010; Fitzpatrick *et al.*, 2015).

Phenotypically, the introduced TULP population resembles a natural LP population in the few traits where this population has been assayed. Magurran *et al*. (1992) found that TULP guppies showed reduced schooling behaviour and increased inspection of a predator compared to TUHP fish, consistent with patterns in other HP/LP comparisons. Fitzpatrick *et al*. (2015) and Gordon *et al*. (2016) found that males from introduced populations evolved more colouration and were larger at maturity, and that introduced females produced larger offspring.

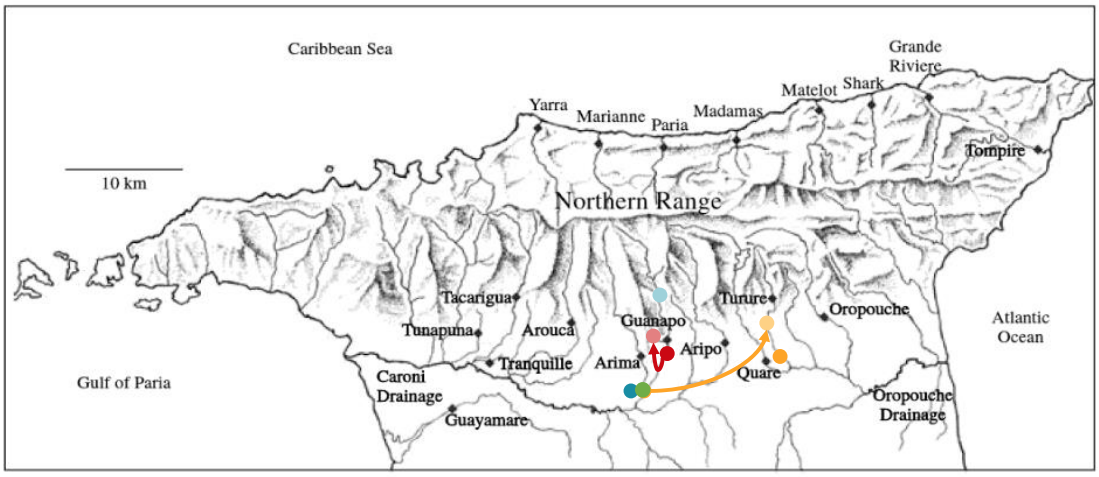


Figure 2.1 Map showing the locations of the sampled populations with arrows indicating direction of the introduction transplants. HP populations in darker colours, LP populations in lighter colours. GLP/GHP are pictured in blue, the ECLP/HP introduction in red. TULP (yellow) was introduced from GHP in green, and TUHP is indicated in dark yellow.

El Cedro River

The second introduction experiment was performed in the El Cedro river by Reznick and Bryga in 1981 (figure 2.1) (1987). The upstream LP site (ECLP) is separated from the downstream HP site (ECHP) by a five-meter waterfall. The profile of ECHP is different from other HP locations: it is smaller, has a lower flow rate and a higher density of guppies compared to other HP sites (Reznick & Endler, 1982), resembling LP sites. Despite this, fish from ECHP have similar life histories to other HP sites (Reznick & Endler, 1982). The introduction involved collecting approximately 100 guppies, including gravid females, and transplanting them to four pools along a 500 meter stretch of the upstream El Cedro river (Reznick & Bryga, 1987). As female guppies are capable of storing sperm, genetic diversity of introduced fish was likely greater than the 100 individuals that were transplanted.

After 10-20 generations, introduced males and females matured at a later age, and males were larger at maturity compared to ECHP fish. Furthermore, females

produced fewer and larger offspring compared to their ECHP counterparts (Reznick & Bryga, 1987). Laboratory-reared second-generation males also showed later age and larger size at maturity compared to their ECHP counterparts, indicating a heritable component to the differences between ECHP and ECLP fish. Laboratory-reared ECLP females on the other hand were not older or larger than ECHP females when they reached maturity. Although, when the laboratory experiment was repeated after 7.5 years (13 generations), females from ECLP matured at a later age and larger size than females from ECHP (Reznick *et al*., 1997). Finally, in 2009, 28 years after the introduction (47 generations), Kemp *et al*. (2009) found males from ECLP evolved smaller colour spots and more iridescence spots compared to HP males.

2.3 Methods

Sampling and data generation

Guppies were sampled from the experimental populations and their sources (ECLP, ECHP, TULP and GHP), from a natural low predation site in the Guanapo river (GLP), and the high predation site in the Turure (TUHP) (figure 2.1). In the spring of 2013, approximately 20 fish from each locality, except from Turure, which was sampled in the spring of 2017 (total N=115, table 2.1). Fish were stored in 95% ethanol at -20°C. DNA was extracted from caudal peduncle tissue using the QIAGEN DNeasy Blood and tissue kit (QIAGEN, Hilden, Germany) following manufacturer's guidelines, with the addition of an on-column RNAase step. Individual samples were prepared for whole genome sequencing following the Illumina TruSeq DNA sample preparation guide with a 250bp insert size. Fish from GHP and GLP were sequenced by multiplexing 6-10 individuals per lane on

an Illumina HiSeq 2000 and 3000 with a 250bp paired-end. The remaining 79 samples were sequenced on an Illumina HiSeq 4000 with 150bp paired end read.

Table 2.1 Summary of population information

Population	ID	Introduction year	N	Males	Females	Mean per individual raw reads	Mean per individual clean reads	Mean depth
Guanapo HP	GHP	N/A	19	9	10	84,794,541	80,947,279	10.4x
Guanapo LP	GLP	N/A	18	10	8	99,093,373	94,135,861	12.1 x
El Cedro HP	ECHP	N/A	19	9	10	59,005,040	58,937,562	8.6 x
El Cedro LP	ECLP	1981	20	10	10	59,333,847	59,261,207	8.8 x
Turure HP	TUHP	N/A	19	9	10	53,693,667	53,624,385	7.2 x
Turure LP	TULP	1956	20	10	10	45,581,608	45,523,708	5.9 x

Read mapping

Prior to processing raw sequence reads, we assessed the quality of the reads with FastQC (Andrews, 2010; Puritz *et al.*, 2014). Adapters and low-quality bases were removed with TRIMGALORE! (Krueger, 2012), followed by running FASTQC a second time to assess pre- and post- quality filtering of the sequencing reads. Trimmed sequence reads were processed following GATK v3.8 and 4.0 Best Practices protocol provided by GATK (v3.8 & 4.0) (van der Auwera *et al.*, 2014). Briefly, reads were mapped to the raw reference genome published in Fraser *et al.* (2020), using BWA-mem 0.7.17 (Li and Durbin, 2009). Duplicate reads were marked and removed using Picard tools v2.17.0 (Broad Institute, 2018) before being merged into a single file per individual. Read quality was recalibrated using variants generated from high-coverage, PCR-free sequencing data (Fraser *et al.*, 2020). Variants were called with HaplotypeCaller in GATK v 4.0, merged with CombineGVCFs and then passed to GenotypeGVCFs, resulting in a combined, genotyped VCF. SNPs were first filtered using a hard quality filter: QualByDepth < 2.0, FisherStrand > 60.0, RMSMappingQuality < 40.0, HaplotypeScore > 13.0, MappingQualityRankSum < -12.5. SNPs were then filtered to retain only biallelic loci with a minimum depth of 5x and a maximum depth of 200x, that were genotyped in at least 50% of all populations, and a minor allele frequency of >0.01. VCFTOOLS (V 0.1.16) (Danecek *et al.*, 2011).

For haplotypes analyses, population VCF files were phased per chromosome using BEAGLE (Browning and Browning, 2007), followed by a second round of phasing using SHAPEIT2 (Delaneau *et al*., 2012). This double-phasing was implemented to optimise the phasing of the VCF files, as per Malinksy *et al*. (2018). SHAPEIT2 has an increased accuracy compared to BEAGLE (Delaneau *et al.*, 2012), but does not accept missing data. Therefore we use BEAGLE, which

does accept missing data but has a high switch rate error (Delaneau *et al*., 2012), to create pre-phased VCF files that can be used as input for SHAPEIT2.

Neutral diversity and population structure

To investigate the evolution of introduction populations, changes in summary statistics between the introduction populations and their HP source population were analysed, as well as between the natural population (GLP) and its HP source GHP. Additionally, we compared TULP to TUHP to investigate HP/LP differences within the Tururе river. Expected heterozygosity (H_e) for each population was calculated with VCFTOOLS (Danecek *et al.*, 2011), and nucleotide diversity (π) and Tajima's D were calculated in PopGenome in R (Pfeifer, 2014).

Population structure was investigated with a principal component analysis (PCA) on all populations using PLINK v1.90. To account for patterns of linkage disequilibrium (LD), the VCF file was pruned for LD by running --indep-pairwise in PLINK (v1.90b6.7, Chang *et al*., 2015) with window size set to 50 bp, step size to 5bp, and the pairwise r^2 threshold to 0.2.

Runs of homozygosity

The distribution and frequency of ROH can provide information on the history of an individual or a population. The presence of short ROH is informative of events further back in time, because ROH from more distant ancestors are broken up into small homozygous segments by recombination over time (Kardos *et al.*, 2017). For example, a population with many short ROH would indicate ancestral family relatedness, which could result from population bottlenecks in the past (Ceballos *et al.*, 2018). The presence of long ROH on the other hand, indicates recent inbreeding, which can be caused by decreased population sizes,

unbalanced paternal contributions and selection (Bertolini *et al.*, 2018). In highly inbred populations, the frequency and total length of ROH are much more variable than in outbred populations; not every individual in the population has consanguineous parents, therefore not every individual will have long ROH (Kirin *et al.*, 2010). The abundance of ROH can also inform on the relative population size: small populations have more ROH compared to large populations as a consequence of increased inbreeding (Ceballos *et al.*, 2018). Finally, admixed populations tend to have fewer ROH than their source population because it brings different haplotypes together (Ceballos *et al*., 2018).

Runs of homozygosity (ROH) were calculated for each individual with a sliding window approach with 50 SNPs per window using PLINK. To minimize the number of ROH that occur by chance, the minimum number of SNPs a ROH must contain was calculated for each population following Purfield *et al*. (2017):

$$l = \frac{\log_e \frac{\alpha}{n_s * n_i}}{\log_e\left(1 - \overline{het}\right)}$$

Where n_s is the number of SNPs per individual (5,169,920 without the sex chromosome), n_i is the number of individuals, α is the percentage of false positive ROH (set to 0.05), and $\overline{\text{het}}$ is the mean heterozygosity across all SNPs. Each run had to be at least 500 kilobases long to exclude short, common ROH present in all individuals and all populations. Finally, at most 1 heterozygous site per window was allowed. Runs of homozygosity were estimated for each individual separately. Resulting ROH were binned into 0.5-0.75 Mb, 0.75-1.0 Mb, 1.0-1.5 Mb, and >1.5 Mb. For each size category, the number of ROH per individual was calculated and used to calculate the population average. Likewise, the total length of ROH was calculated per individual and then averaged per population. Lastly, F_{ROH} was calculated as:

$$F_{ROH} = \frac{\Sigma L_{ROH}}{L_{auto}}$$

Where ΣL_{ROH} is the total length of an individual's ROH above a specified length threshold and L_{auto} is the length of the autosomal genome.

Demographic history analysis in FASTSIMCOAL2

Preparing the SFS

Demographic histories of the populations were reconstructed using the composite-likelihood method in FASTSIMCOAL2 (v2.6) (Excoffier *et al*. 2013). First, folded, 2D site frequency spectra (SFS) were generated for each of the pairwise comparisons: GHP/GLP, ECHP/ECLP, and GHP/TULP. Since we are interested in founding demographic histories, TULP was modelled with GHP as HP counterpart instead of TUHP. We used the minor allele frequency spectrum because of the lack of whole genome data of a suitable outgroup to reliably determine the ancestral state of each allele. To obtain the SFS, we first created a VCF with only monomorphic sites by passing the GVCF cohort to GenotypeGVCFs in GATK v3.8 with the option –includeNonVariantSites to retain monomorphic sites. Monomorphic sites were extracted with the options --selectTypeToExclude INDEL --selectTypeToExclude MIXED --selectTypeToExclude MNP --selectTypeToExclude SYMBOLIC, and --selectTypeToInclude NO_VARIATION, and the same hard quality filters were applied as described above. The filtered monomorphic sites VCF was split by population and the population files were filtered for sites where at least 80% of the individuals were genotyped to minimise missing data. Next, population VCF files containing monomorphic sites were merged to create pairwise files for HP and LP pairs using CombineVariants (GATK, v3.8).

A VCF with SNPs was prepared from the VCF file used for the population genetics analyses, but we applied a more stringent missing data filter (80%). Pairwise VCFs were created for each LP population with its HP source, and were subsequently pruned following the same protocol as described above for the PCA, using PopLDdecay (Zhang *et al*., 2018). The proportion of remaining SNPs was calculated, which was then used to randomly sample the same proportion of the monomorphic VCF in order to maintain the correct SNP to monomorphic sites ratio. From here, the per population SNPs and monomorphic sites VCFs were merged using GATK's (V4.0.4.0) MergeVcfs, and SFS were generated from the pruned VCF files with easySFS (Overcast, 2017) projecting the number of individuals down to minimise missing data in the SFS.

Running the models

Demographic models were fit to the observed 2D SFS of each HP/LP comparison using FASTSIMCOAL2 (Excoffier *et al.*, 2013). The models test whether effective population size (N_E) changed over time, or whether a bottleneck occurred at any time since the split. For GHP/GLP and ECHP/ECLP, the models were tested with and without migration between the populations, while GHP/TULP models did not include migration because the populations reside in different drainages and therefore no migration is possible between the populations. See figure 2.2 for a visual representation of each model, and table 6 for a description of estimated parameters. In model 1, the HP population remained stable, while the LP population was allowed to grow or contract. Similarly, the HP population remained stable in in model 2 and the LP population was subjected to growth/contraction as well as a bottleneck that could occur at any time since the divergence of the populations. In model 3 we allowed population growth or contraction in both the

HP and the LP population, and model 4 included a bottleneck in the LP population. Finally, model 5 included a bottleneck in both populations. The population growth parameter in each model allow for a population to either grow or contract. This was not restricted by the parameters, so based on the SFS, the model estimated population growth or a contraction. In addition, in populations where bottlenecks occurred (models 2, 4 and 5), the bottleneck was set to happen anytime between the origin of the population and the time of sampling. The parameter ranges included in the simulations can be found in supplementary tables S2.1-2.3. Only the initial values were drawn from these ranges, after that the upper boundary does not imply a hard limit for the parameter range and the model can surpass this number by 30% in each cycle (fastsimcoal manual).

Each model was tested by running 100 independent iterations of FASTSIMCOAL2 using the options: -n 200000 (number of simulations), -m (to estimate the SFS for the minor allele), -M (parameter estimation by maximum composite likelihood), and -L 100 (number of ECM cycles). We used a mutation rate of 4.8×10^{-8} (Künstner *et al.*, 2016) and a recombination rate of 4×10^{-5}.

To select the best fitting model per population, the run with the highest likelihood was selected from each model and starting parameters from the best run were used to simulate 100 SFS to obtain a likelihood distribution for each of the models. These distributions were then compared to select the best model. Parameters from the best run of the chosen model were used as initial values to simulate 100 independent SFS per population. By using the parameters from the best run, we can eliminate variance in the optimisation process and instead obtain evolutionary variance. Each simulated SFS was used in two independent runs of FASTSIMCOAL2, using the same settings as above, resulting in 200 sets of

parameter estimates which were then used to calculate medians and the 95% confidence intervals for each parameter.

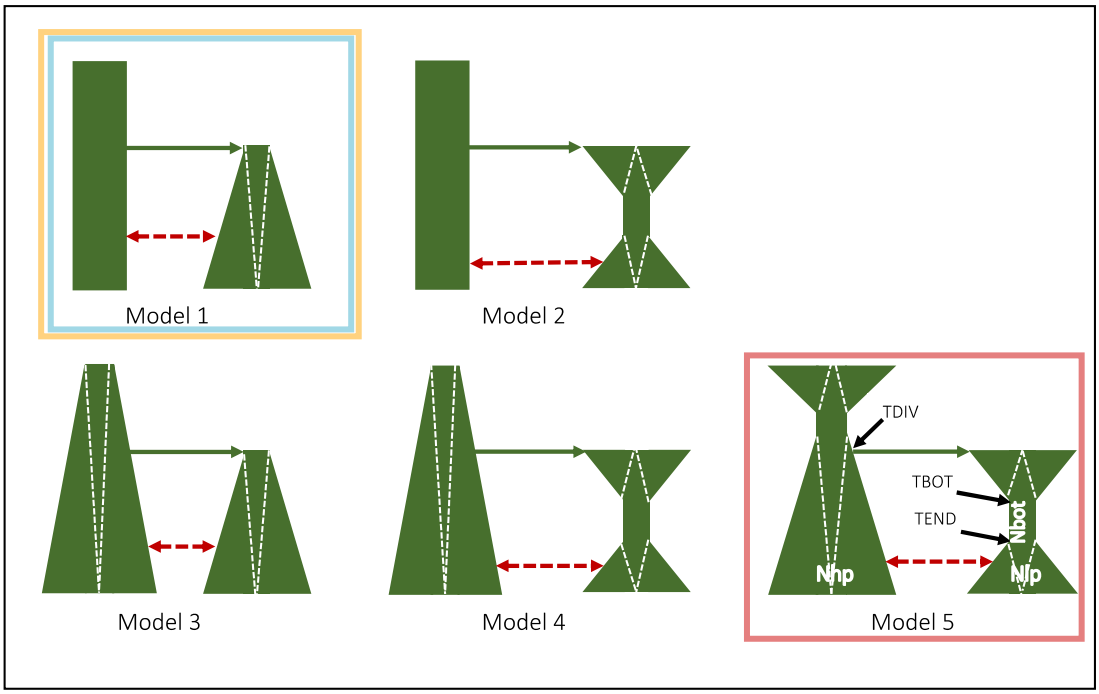


Figure 2.2 Schematics of the demographic models tested with FASTSIMCOAL2. Coloured boxes indicate best fitting models for each population pair: blue = GHP/GLP, red = ECHP/ECLP and yellow = GHP/TULP. The GHP/TULP model did not include migration. Red dashed lines indicate migration, white dashed lines indicate population growth or contraction.

2.4 Results

Across all six populations, whole genome sequencing of 115 individuals resulted in 7.69 billion raw reads, of which 7.52 billion reads remained after filtering (97%). Average population coverage was 8.8x and ranged from 5.9x (TULP) to 12.1x (GLP) (table 2.1). The final processed Variant Call Format (VCF) contained 5,389,971 SNPs.

Population structure

In a PCA, the first axis (PC1) captured 18.95% of the variation among populations and separates the different rivers from each other. While, PC2 explains 16.63% of variation and separates GLP from the other populations and

the Turure river from the Guanapo and El Cedro rivers (figure 2.3). Finally, PC3 explains only 6.14% of variation, and separates the LP populations from the HP population within a river. Pairwise F_{ST} revealed similar patterns: comparisons involving GLP were the highest compared to all other population comparisons and TULP, TUHP, and GHP were least differentiated (table 2.2). In addition to the whole genome set, we investigated global F_{ST} values of the data set with outlier windows removed (supplementary table S2.4) and with the sex chromosome (chromosome 12) removed (supplementary table S2.5). Removing outlier windows resulted in a very minor, but significant, reduction of F_{ST} values of an average 0.009 for the median and 0.016 for the mean. However, it made no difference for the observed patterns among population pairs described above. Removing the sex chromosome resulted in minor changes in ten of the fifteen pairwise comparisons, but none of these were significant ($p>0.4$ for all comparisons, Mann-Whitney U test).

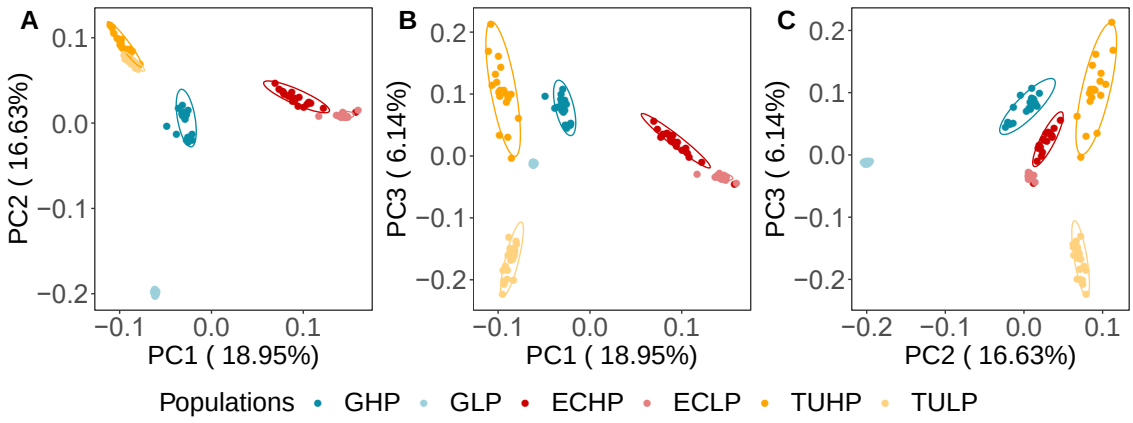


Figure 2.3 Principal component analysis of all six populations: A) Plotting PC1 (18.95% of variance explained) against PC2 (16.63% of variance explained). B) PC1 and PC3 (6.14% of variance explained), C) PC2 and PC3. Ellipses indicate the 95% confidence interval.

Table 2.2 Pairwise global F_{ST} values for each population. Mean F_{ST} below the diagonal, median F_{ST} above the diagonal. Population names can be found in table 2.1.

	GHP	GLP	ECHP	ECLP	TUHP	TULP
GHP		0.292	0.091	0.202	0.036	0.056
GLP	0.302		0.467	0.631	0.296	0.336
ECHP	0.119	0.458		0.061	0.120	0.151
ECLP	0.231	0.596	0.092		0.215	0.257
TUHP	0.046	0.311	0.141	0.238		0.032
TULP	0.073	0.350	0.175	0.280	0.041	

Neutral diversity

In general, median expected heterozygosity (H_e) and nucleotide diversity (π) were significantly lower in LP populations than their HP counterparts for all within-river comparisons. GLP had the lowest values of H_e (0.006) and π (0.00004), a reduction of 96.4% and 96.8%, respectively, from GHP values ($p<0.0001$, Mann–Whitney U test, table 2.3, figure 2.4 left column). In contrast, after the introduction from GHP into TULP, H_e and π increased significantly ($p=1.5e-11$ and $p=1.2e-10$ respectively, Mann–Whitney U test, table 2.3 and figure 2.4 third column), however the difference was not large, with a median difference of 6.5% and 8.1%, respectively. Finally, in ECHP/ECLP we found low values of H_e and π in both populations, the transplant from ECHP to ECLP further reduced these values in ECLP: median H_e decreased by 56.5% (from 0.124 to 0.054), and median π by 58.4% (from 0.00089 to 0.00037, $p<0.0001$, Mann–Whitney U test, table 2.3, figure 4 second column).

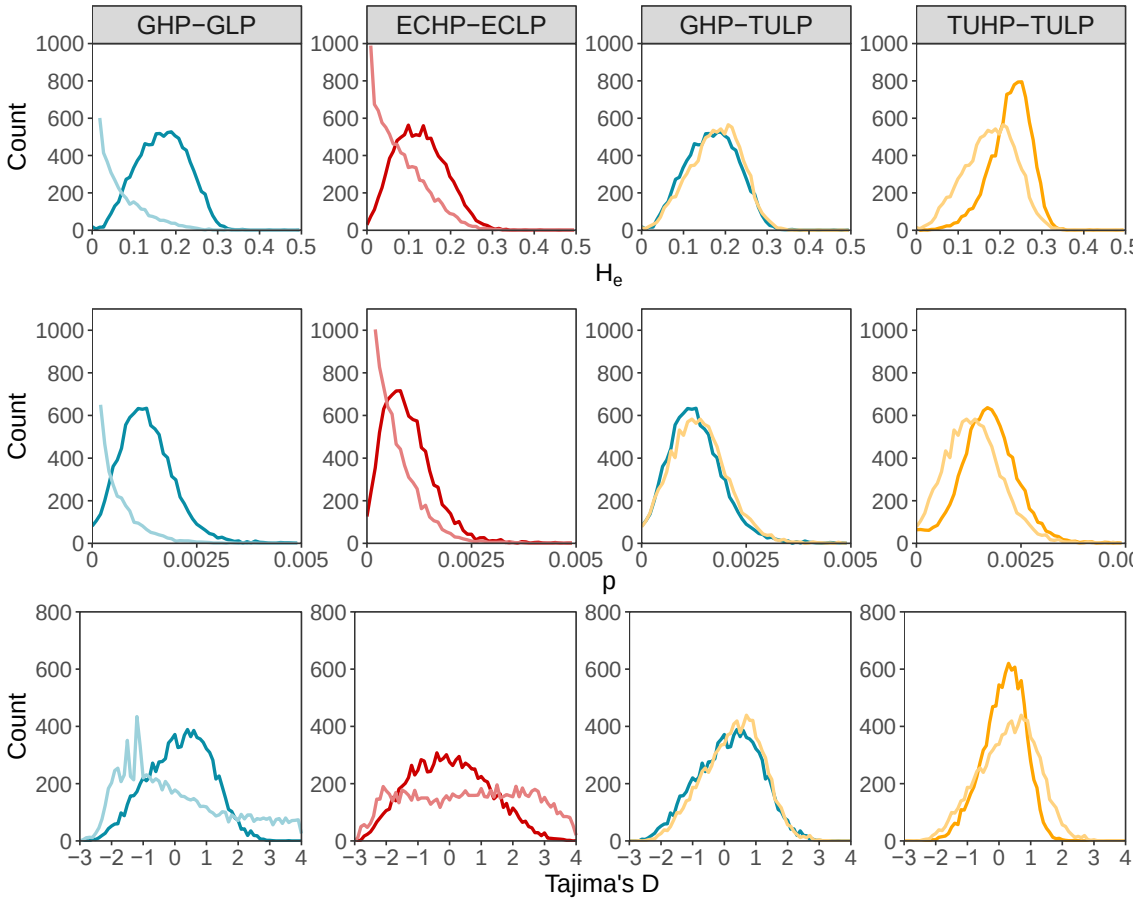


Figure 2.4 Frequency distributions of summary statistics, expected heterozygosity (He, top panel), nucleotide diversity (π, middle panel) and Tajima's D (bottom panel).

Table 2.3 Summary of population genetic statistics. Shown are mean, median and 95% confidence boundaries for expected heterozygosity (H_e), and mean, median and minimum and maximum values for nucleotide diversity (π) and Tajima's D (TD).

Population	Mean H_e	Median H_e	95% CI	mean π	median π	min-max π	mean TD	median TD	min-max TD
GHP	0.168	0.169	0.053 - 0.280	0.0013	0.00124	0 – 0.008	0.161	0.224	-2.641 – 3.551
GLP	0.030	0.006	0 - 0.178	0.0002	0.00004	0 – 0.005	0.022	-0.385	-2.864 – 4.119
ECHP	0.127	0.124	0.024 - 0.250	0.0010	0.00089	0 – 0.007	0.009	-0.046	-2.824 – 3.974
ECLP	0.069	0.054	0.001 - 0.214	0.0005	0.00037	0 – 0.007	0.580	0.626	-2.903 – 4.122
TUHP	0.225	0.230	0.112 - 0.303	0.0017	0.00172	0 – 0.008	0.156	0.215	-2.393 – 2.495
TULP	0.177	0.180	0.055 - 0.285	0.0014	0.00134	0 – 0.007	0.290	0.357	-2.704 – 3.258

The distribution of Tajima's D was shifted towards a more positive median in the introduced populations compared to their HP counterparts ($p<0.0001$, Mann–Whitney U test, figure 2.3 bottom panel), indicating a loss of rare alleles relative to common alleles, a pattern observed after a sudden population contraction. In GLP and ECHP, Tajima's D was negative, the result of an excess of rare alleles which suggests a population expansion after a bottleneck.

Runs of homozygosity

We identified ROHs of > 0.5 Mb in 108 of the 115 sequenced individuals (one individual in GHP and six individuals in TUHP lacked ROH >0.5 Mb). The frequency of ROH varied greatly among the populations, where GHP and TUHP had very few (<100), and GLP and ECLP had many ROH (>2500) (table 2.4). ROH also ranged in size among populations: ROH > 2Mb can only be found in GLP, ECHP and ECLP (maximum length was 2.80 Mb, 2.09 Mb and 2.3 Mb, respectively), whereas the maximum length in TULP is 1.62 Mb, and both GHP and TUHP lack ROH > 1.5 Mb (table 2.4, figure 2.5). GLP and ECLP (which had the lowest values of H_e and π), have many more ROH in all size classes and they have more ROH > 1.5Mb than the other populations, suggesting a history of bottlenecks and inbreeding in these populations (table 2.4, figure 2.5). Likewise, ECHP has more ROH in every size class, but fewer than GLP and ECLP. The observed patterns of low abundance and a low total length of ROH in GHP, TUHP and TULP suggest these populations are large and did not experience bottlenecks or inbreeding in their pasts. Furthermore, the extremely low number of ROH in GHP and TUHP can also be indicative of admixture from other populations further upstream in their respective rivers.

Table 2.4 The average length and number of ROH per individual in each population.

Population	Total # ROH	Average number	Min-max length (Mb)	F_{ROH} (min-max)
GHP	88	5	0.50-1.25	0.004 (0.0007-0.013)
GLP	4325	240	0.50-2.80	0.243 (0.116-0.313)
ECHP	719	38	0.50-2.09	0.035 (0.013-0.081)
ECLP	2691	135	0.50-2.38	0.133 (0.036-0.298)
TUHP	51	4	0.50-1.08	0.004 (0.0007-0.010)
TULP	293	15	0.50-1.62	0.013 (0.005-0.043)

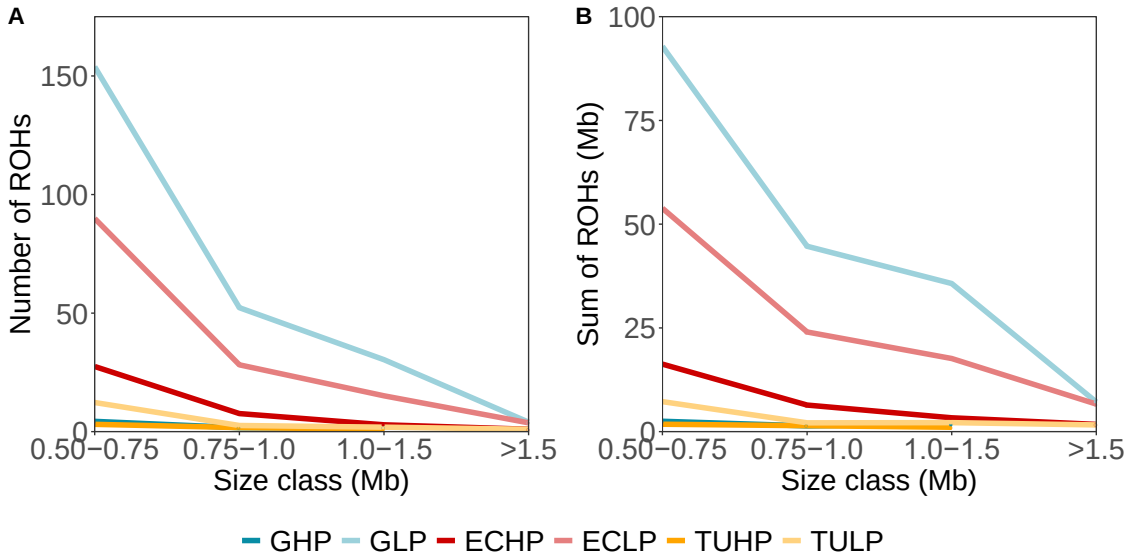


Figure 2.5 Plots showing ROH per size category. A) Mean number of ROH's per individual for each size category. B) Mean sum of ROH's per individual for each size category.

We also calculated the inbreeding coefficient, F_{ROH}, as the proportion of the autosomes covered in ROH. GLP and ECLP had significantly higher values of F_{ROH} than the other populations ($p<0.0001$, Mann–Whitney U test, table 2.4, figure 2.6), and they had large variability in genome coverage among individuals, ranging from 0.116 to 0.313 in GLP and from 0.035 to 0.298 in ECLP, meaning that some of the individuals have close to a third of their genome covered in ROH. This is consistent with the idea that these populations have been subjected to bottlenecks and inbreeding. In ECHP F_{ROH} ranged from 0.013 to 0.081, with an

average of 0.035. Even though this is significantly less than GLP and ECLP ($p<0.0001$, Mann–Whitney U test), it is nearly triple the value of TULP (0.013) and almost nine times as much as GHP and TUHP (0.004 for both, $p<0.0001$, Mann-Whitney U test) (table 2.4, figure 2.6), suggesting ECHP also experienced inbreeding in the past.

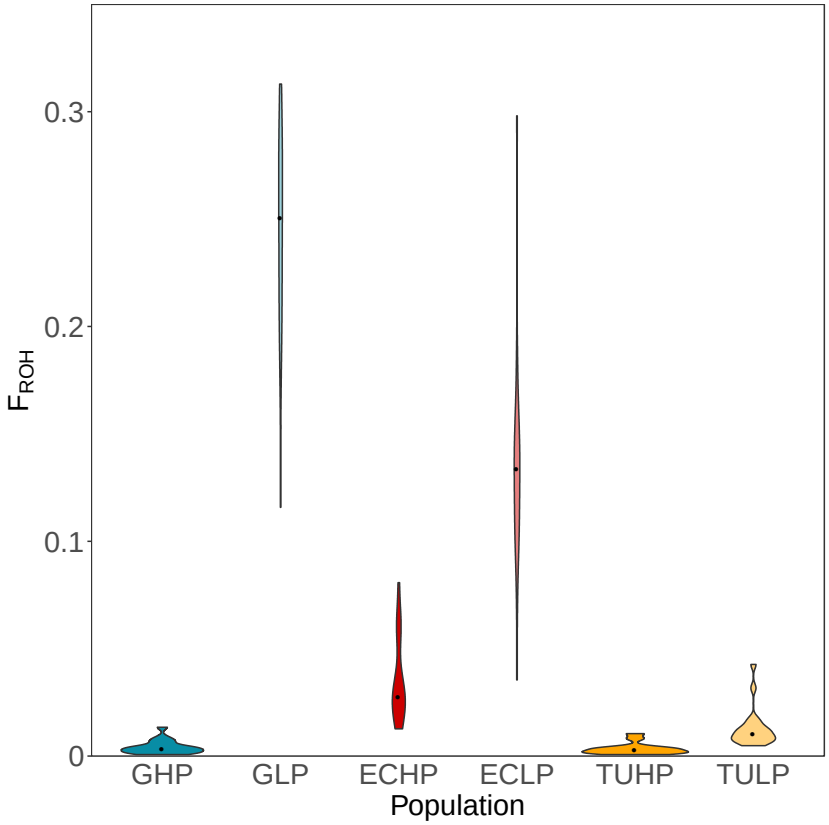


Figure 2.6 F_{ROH} per individual per population. F_{ROH} is the proportion of the genome covered in ROH.

Demographic modelling

Across the five models, model 1 had the highest likelihood in GHP/GLP and in GHP/TULP, meaning no bottlenecks occurred in any of the three populations, and in the case of GHP/GLP, migration was likely between the populations (table 2.5, figure 2.7A/B & E/F). Figure 2.7F shows that model 1 and model 3 have

nearly equal likelihood distributions in GHP/TULP. The two models differ in growth for GHP (figure 2.2), however the value of growth for GHP in model 3 is almost negligible (1.0E-07 individual/generation), so we used model 1 for further analysis of the GHP/TULP comparison. In ECHP and ECLP, model 5M had the best fit, with a bottleneck in both populations and migration between the two populations (table 2.5, figure 2.7C/D).

Table 2.5 Δ-likelihoods for the best runs of each model in FASTSIMCOAL2. These runs were then used to simulate a likelihood distribution of 100 SFS. The best models are printed in bold

	No migration					Migration				
Population	**model 1**	**model 2**	**model 3**	**model 4**	**model 5**	**model 1**	**model 2**	**model 3**	**model 4**	**model 5**
GHP-GLP	-71.9	-26356.4	-71.6	-32796.6	-51070.1	**-48.5**	-160.1	-160.4	-163.9	-82.0
ECHP-ECLP	103.8	-103.7	-103.9	-103.7	-103.6	-29.7	-22.2	-29.9	-18.6	**-17.0**
GHP-TULP	**-87.1**	-100.1	-87.7	-100.0	-101.4	NA	NA	NA	NA	NA

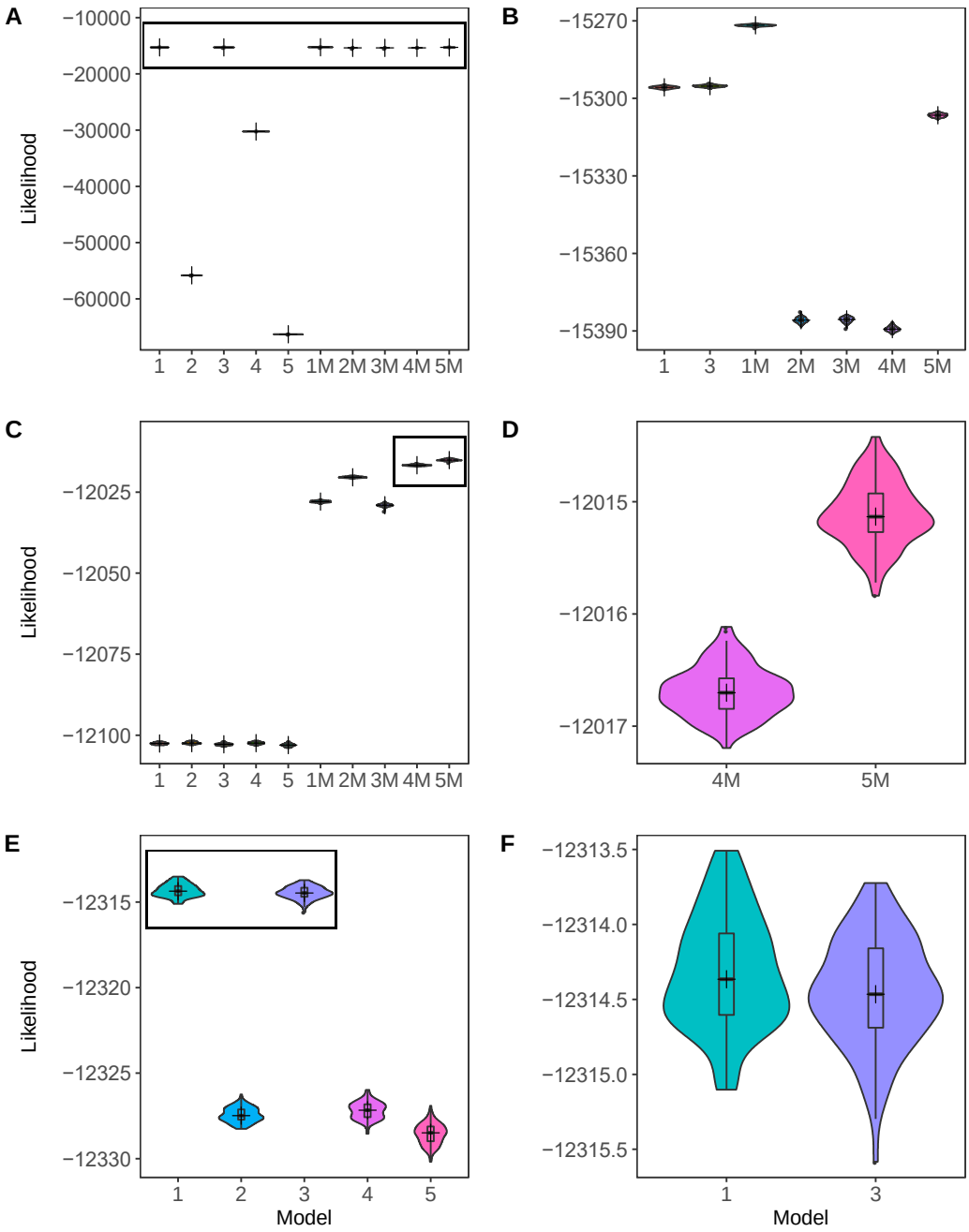


Figure 2.7 Likelihood distributions of each model for each HP-LP pair, obtained from the best initial runs in FASTSIMCOAL2. A) GHP/GLP models, C) ECHP/ECLP models, and D) GHP/TULP models. Panels on the righthand side (B, D, F) are a close-up of the black boxes in A, C and E.

Population parameters were estimated by simulating the best model, using the parameters of the best run as initial values. Migration was likely between GHP and GLP, and ECHP and ECLP, with a higher probability of individuals migrating downstream into the HP populations than vice versa (table 2.6, figure 2.8). Population splits for both introduced populations were estimated to be slightly

older than the actual times of introduction: estimated divergence time in ECHP/ECLP was 92 generations (the introduction was 64 generations ago, assuming a generation time of 2 generations per year (Endler, 1980)), while estimated divergence time for GHP/TULP was 193 generations ago (the introduction was 114 generations ago) For GLP, the model estimated time since divergence from GHP at 315 generations ago.

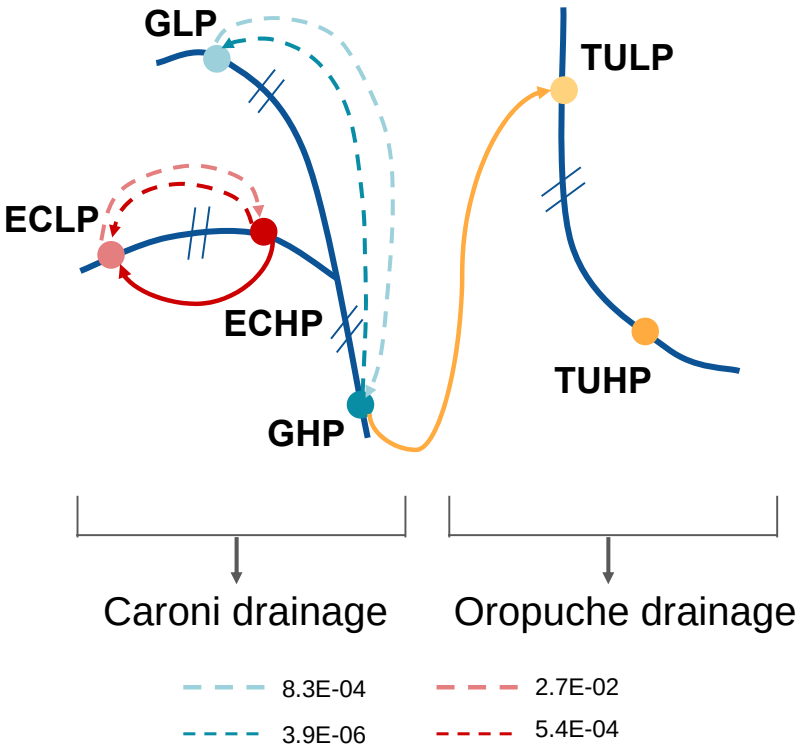


Figure 2.8 Schematic showing migration between ECHP/ECLP and between GHP/GLP. Migration is given in the probability of a gene moving to another population. Downstream migration has a higher probability of occurring than upstream migration. Migration from ECLP into ECHP is more likely than migration from GLP into GHP.

Simulated estimates for current population size suggest a large N_E for GHP (based on GHP/TULP model) and TULP (N=37,343 and N=29,161, respectively, 95% CI's can be found in table 2.6) and a much smaller N_E for GLP and ECLP (78 and 333, respectively, 95% CI's in table 2.6), consistent with genetic diversity estimates and patterns observed in the ROH analysis (tables 2.3 & 2.4). Ancestral effective population sizes for the introduced ECLP population was nearly six

times larger than the estimate for TULP (N=2,084, and N=337, respectively, 95% CI's in table 2.6). The ancestral population size of GLP was much smaller (N=103, 95% CI 2-7423), however, the confidence interval was large, indicating the model struggled to converge on a consistent estimate.

The ECLP/ECHP best fitting model was the only best fitting model with bottlenecks. ECLP shrank until the start of the bottleneck, estimated to begin 18 (95% CI 13-22) generations ago. The bottleneck ended 14 generations ago (95% CI 9-18), during which the population size collapsed to N=32 (95% CI 27-38). Since then, ECLP has recovered to an estimated effective population size of 333 (table 2.6, figure 2.9). The bottleneck in ECHP was estimated to start 44.2 million generations ago (22 million years) and lasted for approximately 22 million generations (table 6), during which it collapsed to an effective population size of N=114. Since then, it has grown and reached a current N_E of N=10,675.

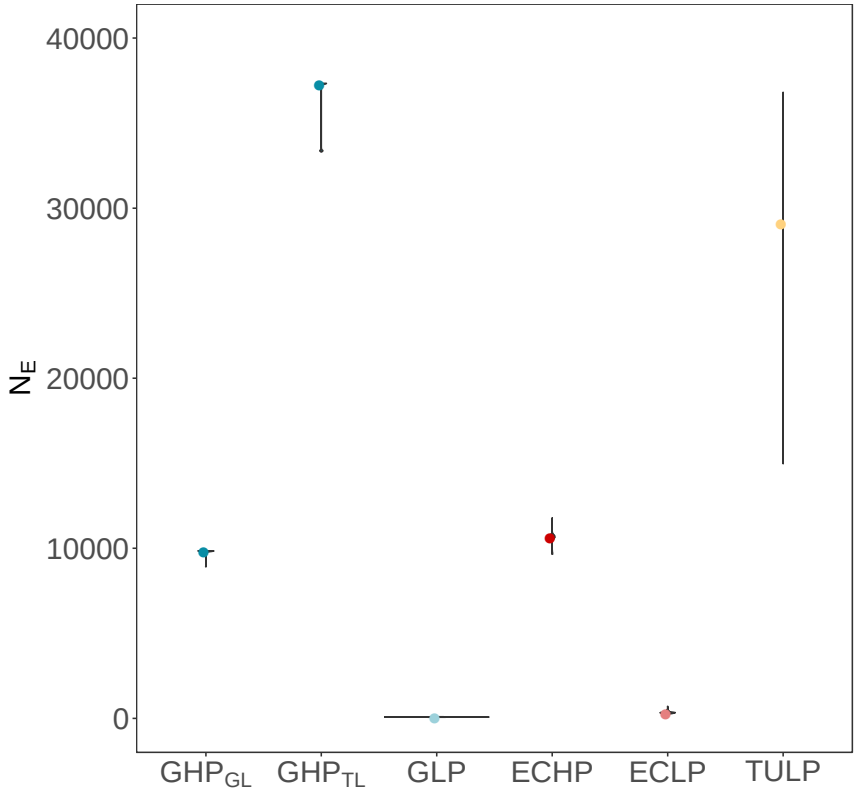


Figure 2.9 Parameter estimates of current effective populations size, based on simulated SFS in FASTSIMCOAL2. GHP_{GL} and GHP_{TL} indicate the estimate for GHP in the GLP and TULP model, respectively. Coloured points are medians of 200 simulated runs.

Table 2.6 FASTSIMCOAL2 parameter estimates from the best run and 95% confidence intervals generated by simulations of the best model in ILL. Population size estimates are given in diploid number of animals, and times are given in generations. Migration rates are the probability of a gene to move from one population to other per generation. Growth rates are given per generation, positive values indicate population contraction forward in time. BN = bottleneck. Values marked with * indicate estimates that do not fit within the simulated 95% confidence interval. (table on the next page)

A) GHP/GLP

Parameter	Median	95% CI	
		Lower bound	Upper bound
GHP N_E	9849	9796	9849
GLP N_E	78	68	92
GLP ancestral N_E	103	2	7423
GHP/GLP divergence	315	280	384
GLP growth	8.90E-09	-9.50E-08	1.40E-07
GLP>GHP migration	8.30E-04	6.70E-04	1.00E-03
GHP>GLP migration	3.90E-06	5.30E-08	7.00E-05

B) ECHP/ECLP model

Parameter	Median	95% CI	
		Lower bound	Upper bound
ECHP N_E	10675	9662	10675
ECLP N_E	333	237	488
ECLP ancestral N_E	2084	906	3756
ECHP BN N_E	114	2	6304
ECLP BN N_E	32	27	38
ECHP/ECLP divergence	92	57	116
ECHP growth	-2.20E-07	-7.40E-07	-2.10E-08
ECLP growth	-1.50E-01	-2.40E-01	-1.20E-01
ECHP growth before BN	1.80E-08	-8.70E-07	1.00E-06
ECLP growth before BN	5.50E-02	4.50E-02	9.70E-02
ECLP>ECHP migration	2.70E-02	2.20E-02	3.60E-02
ECHP>ECLP migration	5.40E-04	3.30E-06	1.10E-03
ECLP BN start	18	13	22
ECLP BN end	14	9	18
ECHP BN start	4.40E+07	3.40E+07	5.30E+07
ECHP BN end	2.10E+07	6.80E+06	3.90E+07

C) GHP/TULP model

Parameter	Median	95% CI	
		Lower bound	Upper bound
GHP N_E	37343	33378	37343
TULP N_E	29161	19330	36814
TULP ancestral N_E	337	243	509
GHP/TULP divergence	193	183	197
TULP growth	-3.00E-02	-3.30E-02	-2.70E-02

2.5 Discussion

Here, we used patterns of neutral genomic variation and inferred demographic histories to investigate evolution in six populations of guppies, including two experimental populations. Demographic processes can leave traces in the genome resembling those left behind by selective events, and they can influence the course of natural selection (Tajima, 1989b; Galtier *et al.*, 2000; Rosenblum *et al.*, 2014), so it is important to understand the demographic history of a population before examining patterns of selection. Our data uncover varying demographic histories among the six populations. We found evidence of bottlenecks and inbreeding in GLP, ECHP and ECLP, but not in GHP, TUHP and TULP. The most extreme signals were found in the natural LP population (GLP), which had the lowest genetic variation and the smallest population size. The signals in ECHP were weaker, but our demographic model showed it experienced a bottleneck in the past, and values of expected heterozygosity and nucleotide diversity were low compared to GHP, TUHP and TULP. Introduction to a new environment can affect a population in different ways: some translocations result in a loss of genetic variation, whereas others do not seem to affect the population as much (Lambert *et al.*, 2005). We found this to be the case in our two introduced populations: ECLP experienced a bottleneck that drastically reduced its population size and genetic variation, whereas TULP experienced extensive population growth and saw a slight increase of genetic variation since the introduction.

In the El Cedro river, we found low levels of nucleotide diversity and expected heterozygosity in the ECHP source population, and the introduction resulted in even lower values in ECLP. This difference in diversity between HP and LP populations is consistent with previous research (Barson *et al*., 2009; Crispo *et*

al., 2006; Fraser *et al*., 2015; Suk & Neff, 2009; Willing *et al*., 2010), and in studies of other species where populations are separated by physical barriers (Hamilton *et al*., 1989; Hughes *et al*., 1995; Castric *et al*., 2001; Guy *et al*., 2008). Our demographic analysis revealed ECHP and ECLP both experienced a bottleneck in their past. In ECLP we found a mean positive Tajima's D, compared to a mean negative Tajima's D in ECHP, suggesting a loss of rare alleles, concurrent with a sudden population contraction like a founding bottleneck. A similar change was observed in a natural introduction experiment with stickleback (Marques *et al*., 2018). The proportion of the genome covered by ROH (F_{ROH}) in both populations was high compared to the other populations (except GLP). F_{ROH} in ECLP (0.133) is comparable to F_{ROH} observed in highly inbred human populations (Karitiana F_{ROH}=0.15) (Prado-Martinez *et al.*, 2013), and F_{ROH} in ECHP is lower (0.035), but still considered inbred compared to wild populations of other species. For example, F_{ROH} in most flycatcher populations is < 0.03 (Kardos *et al.*, 2017), and Hooper *et al*. (2020) considered populations of killer whales inbred when F_{ROH} > 0.015. F_{ROH} was nearly four times lower in ECHP compared to ECLP, and the number and length of ROH was much lower, which suggests the bottleneck in ECHP was less intense. Our demographic model estimated ECHP's current effective population size to be over thirty times larger than ECLP, consistent with patterns found in Barson *et al*. (2009) and Fraser *et al*. (2015): LP populations have lower or similar population sizes compared to their HP source. Such patterns are often observed in riverine species (Hughes *et al*., 1995; Guy *et al*., 2008), and lower estimates of N_E are especially found in populations inhabiting upstream rivers in mountainous areas (Cook *et al*., 2011). The evidence of a population bottleneck together with a large current population size in ECHP fits with the negative value of Tajima's D observed in this population: an excess of

rare alleles can result from a population expansion after a bottleneck. Congruent with previous work (Fitzpatrick *et al*., 2015; Fraser *et al*., 2015), population differentiation was low between the two El Cedro populations. Contrary to Fraser *et al*. (2015) however, we found evidence for migration between ECHP and ECLP in our demographic model, with a higher probability of downstream migration than upstream. This migration could account for the low levels of differentiation between these populations.

A different story emerges from the older introduction experiment in the Tururе. Here, levels of H_e and π in TULP were similar, even slightly increased, compared to those in its HP source, GHP, and they were much higher than the values observed in ECHP and ECLP. Similar to ECLP though, we found Tajima's D had shifted towards a more positive value in TULP compared to both GHP and TUHP, suggesting a founding bottleneck. A previous study of the TULP introduction, using 25 allozyme loci also found evidence for a founding bottleneck in TULP (Shaw *et al.*, 1992). However, the high levels of nucleotide diversity and expected heterozygosity suggest population expansion after the introduction may have been rapid enough to avoid initial inbreeding and pervasive genetic drift. Furthermore, TULP lacked a strong bottleneck/inbreeding signature in the ROH analysis and N_E was much higher in TULP compared to populations where we found evidence of (founding) bottlenecks and inbreeding (figure 9). Like TULP, GHP had fewer and shorter ROH, and it lacked any ROH > 1.5 Mb, indicating a large, stable population, which likely has not experienced bottleneck and inbreeding events. This was further confirmed by our best fit demographic model which did not include a bottleneck or population growth and estimated current effective population size of GHP to be larger than any of the other populations.

In addition to comparing TULP to GHP, we compared TULP to an HP population within the same river (TUHP). Other studies have found considerable gene flow from TULP into Turure HP populations, even reaching the most downstream populations like the HP population considered in this study (TUHP) (Suk and Neff, 2009; Fitzpatrick *et al.*, 2015). In agreement with these studies, our ROH analysis found very few ROH in TUHP, suggesting it experiences admixture from other populations. Furthermore, we found low levels of differentiation between TUHP and TULP which could be explained by extensive gene flow between these populations. Gene flow may explain the surprisingly low F_{ST} between GHP and TUHP: even though they originate from two different drainages (Caroni and Oropuche); if gene flow from TULP (an exclusively Caroni population) into TUHP is high enough, it could have changed the genetic footprint in TUHP. Congruently, Russell (2004, in Magurran, 2005) found over 90% of the individuals of TUHP populations had introduced nuclear alleles from the Caroni drainage.

Finally, we included a naturally colonised LP population within the study river for comparisons, GLP. GLP was genetically more differentiated from the other populations and had significantly lower values of H_e and π compared to GHP. Indeed, the values for GLP were the lowest among all populations studied. In combination with more and longer ROH in this population and a high F_{ROH}, this indicates GLP is a small population, strongly affected by historical bottlenecks and inbreeding. Our demographic analysis further revealed that the founding N_E of GLP was likely extremely small compared to the other LP populations, suggesting that the founding bottleneck was more intense. This low founding N_E is corresponds with the idea that natural LP populations can be colonised by very few individuals (Fraser *et al*., 2015). According to our demographic model, GLP

remained small and is currently still the smallest population. The model also revealed migration between the populations, with more migration downstream than upstream, however the levels of migration were much lower than those observed between ECHP and ECLP.

Our demographic models for GLP and TULP share the same HP population. However, estimates for current N_E of GHP in the GHP/GLP model and the GHP/TULP model were disparate (N=9,849 vs N=37,343, respectively). One explanation could be that the GLP/GHP model is not very accurate due to the high level of inbreeding in GLP. Coalescence-based methods that generate SFS from a demographic model, assume random mating within a population, which might not be true for a highly inbred population like GLP. The increased level of homozygosity in an inbred population can lead to changes in the observed SFS that cannot be detected by models that assume panmixia, which may subsequently affect estimates of demography (Blischak *et al*., 2020). Blischak *et al*. (2020) used simulations to show that in two-populations models, where one population is inbred, parameter estimates were underestimated, as seems to be the case for the GHP estimate from the GLP model compared to the estimate from the TULP model.

Unreliable parameter estimates could explain additional aspects of the GHP/GLP model that do not correspond with other results in this study, or what is generally known about these natural populations. For example, divergence time is estimated to be 315 generations, which is younger than might be expected considering it is a naturally colonised population. Alternatively, this seemingly young divergence time could be explained if GLP has experienced several colonisation/extinction events, during which it was recolonised by just a few individuals. Seasonal flooding is known to have a severe impact on the population

size of upstream locations, and can even result in local extinctions in guppies (Grether *et al*., 2001). If the population in GLP has gone extinct and was subsequently recolonised several times, it might be older than was estimated here, and it would be impossible to distinguish between changes in population size and rates of extinction/recolonization (Wakeley and Aliacar, 2001). We also found reduced effective population size, reduced nucleotide diversity, and increased values of F_{ST}, all characteristics that can result from multiple founder events (Pannell, 2003; Lambert *et al.*, 2005).

The example above illustrates why it is useful to use multiple methods to infer demographic pasts of populations. By comparing and combining the results of multiple analyses we were able to more confidently reconstruct the demographic histories of the populations under investigation. Furthermore, the different analyses complement each other: with our sample sizes, methods based on the SFS are generally more useful for identifying demographic events further in the past, whereas methods based on linkage disequilibrium and recombination, such as our ROH analysis, can also detect more recent events (Depaulis *et al.*, 2003; Beichman *et al.*, 2018).

Overall, we found varying demographic histories among experimental populations of guppies. Accounting for demographic variation is essential when looking for selection, especially when the goal is to find instances of molecular convergent evolution. As we have seen in this chapter, demographic events can hugely affect the amount of genetic variation available in a population. This in turn, affects the efficiency of selection and the probability of convergent evolution among populations (Rosenblum *et al*., 2014), where a larger proportion of shared ancestral variation increases the probability of molecular convergence when adapting to a new environment (Barrett and Schluter, 2008). In the next chapter,

we search for signatures of selection along the genome in these experiments and investigate if the phenotypic convergence is underpinned by convergence at the molecular level. These two experiments represent two very different scenarios, one that is older and likely started with high genetic diversity, the other a younger, smaller population, where the source population itself had reduced genetic variation. Using the information about the various demographic parameters inferred in this chapter, we will account for these differences in demography when looking for signals of selection.

Chapter 3 – Effects of starting genetic variation on convergent evolution in established experimental populations of guppies

3.1 Abstract

Differences in demographic histories between populations could lead to differences in available genetic variation, which in turn can lead to differences in selection and the probability of molecular convergence underlying convergent phenotypes. In the previous chapter, we found that two experimentally introduced populations of guppies represent different scenarios, one older population that originated from a source with high levels of genetic variation and the other a younger population originating from low levels of genetic variation. In this chapter we investigate how these differences in founding genetic variation have affected selection within a river, and how it affected the occurrence of molecular convergence among rivers. Using whole-genome sequencing data we found population specific signals of selection in all population pairs, despite low levels of standing genetic variation in two of the three populations. We also show that both introduced populations show molecular convergence with a natural low predation population, but we found little evidence for convergence between the two experimental populations, suggesting diverging molecular pathways to adapt to low predation environments.

3.2 Introduction

Convergent evolution, where independent populations evolve similar phenotypes under similar environmental pressures, can reflect both the strength and limitations of selection. Convergence of phenotypes can arise through

changes in different genes and pathways resulting in the same phenotype (e.g. Nachman *et al*., 2003; Roelants *et al*., 2010), or from underlying molecular convergence (Andreev *et al*., 1999; Colosimo *et al*., 2005). Molecular convergence can be achieved via different mechanisms: 1) selection acting on a new (*de novo)* mutation (DNM) arising independently in different populations, 2) a locus being shared from one population into another by gene flow, or 3) selection on a genetic variant that was polymorphic in a common ancestor (standing genetic variation, SGV) (Stern, 2013; Lee and Coop, 2019). In some cases, such as recently established populations, the role of SGV is possibly more important than evolution through DNMs (MacPherson and Nuismer, 2017). Here, adaptation is more likely to occur through SGV as the beneficial allele is already present in the population and usually exists at higher frequencies than a new mutation (Barrett and Schluter, 2008). Even though adaptation via SGV is expected to increase molecular convergence in the early stages of adaptation, it can also be the cause of limited convergence if populations harbour different amounts of SGV, and consequently a different set of variants to adapt from.

The amount of SGV in a population is affected by its demographic history, therefore differences in demography between populations can generate historical contingencies that in turn result in converging/diverging outcomes of evolution (Simões *et al.*, 2017). Events affecting population sizes, such as bottlenecks, can play an important role in the determination of selection outcomes. Population bottlenecks can reduce genetic variation (Mayr, 1963), which in turn can reduce the adaptive potential of a population and reduce the likelihood of convergence through SGV. Small population sizes also lead to a reduced efficacy of selection to remove deleterious alleles and fix advantageous alleles resulting in a greater

impact of random genetic drift (Charlesworth, 2009) and a reduced likelihood of convergent evolution at the molecular level (Lachapelle *et al*., 2015).

The experimental evolution framework offers a powerful tool to study the effects of genetic variation on the outcome of selection. Experimental evolution allows for control over the environments, and in combination with whole genome sequencing, allows us to track allele changes in real-time (Stern, 2013). Studies have shown various examples of convergent genomic evolution in *Drosophila* (e.g. Graves Jr *et al.*, 2017) and bacteria (Liao *et al*., 1986; Tenaillon *et al*., 2016), as well as fish (e.g. Jones *et al.*, 2012) and plants (Christin *et al.*, 2007), although molecular convergence is not always the case. For example, Wittkopp *et al*. (2003) found that phenotypic convergence of pigmentation patterns in various *Drosophila* species result from underlying changes in different genes. In studies specifically investigating the role of SGV and the likelihood of molecular convergence, variation in SGV is often achieved by changing population sizes at the start of the experiment. Wein and Dagan (2019) investigated the effect of bottleneck size on genetic diversity and evolvability of bacteria and observed highest genetic variation and an increasing degree of molecular convergence in larger populations. In contrast, Vogwill *et al*. (2016) found that convergence was more likely under weak, as well as, intense bottlenecking, but that under intermediate bottlenecks resistance to antibiotics evolved through a greater diversity of genetic mechanisms, resulting in a reduced level of genetic convergence.

A drawback of most experimental evolution studies is that they are often performed with model species or laboratory adapted strains, and therefore do not provide an insight in the process of adaptation in a natural environment (Wood *et al*., 2005). Studies that use individuals from the wild and subsequently perform

experiments in a laboratory setting (Matos *et al.*, 2015), or in semi-natural ponds (Rennison *et al*., 2019), can address some of these biases but even these studies do not provide the complex structure of a natural environment. Experimental evolution studies with natural populations in the wild investigating convergent evolution are rare, most likely because of the complexity of natural populations (for example logistical, ethical and conservational challenges) (Kawecki *et al.*, 2012). Studying wild experimental populations will allow researchers to examine the effects of the complex natural environment on early adaptation to a new environment, and investigate the likelihood of convergent evolution among populations with different evolutionary histories (Fraser and Whiting, 2019).

The Trinidadian guppy: introduction experiments in the wild

Several long-term experimental populations of guppies in Northern Range Mountains of Trinidad provide an excellent system to study convergent evolution in the wild. Here, guppies from high predation (HP) locations were transplanted to previously guppy-free low predation (LP) locations, and their phenotypic evolution was followed throughout the years. This chapter focuses on two such introduction experiments (for a detailed description of each experiment, see the introduction of Chapter 2). Briefly, the first introduction into the Turure river was performed in 1956, when circa 200 guppies were translocated from a HP location in the Guanapo river (GHP, Caroni drainage) to a, previously guppy free, LP location (TULP, Oropuche drainage). The introduced guppies have since invaded several downstream HP environments, leaving a genetic trace of the Caroni drainage (e.g. Shaw *et al*., 1992; Suk & Neff, 2009). The second experiment was performed in the El Cedro river in 1981. Here, 100 guppies, including gravid females, were moved from a small HP location (ECHP) to an LP location (ECLP)

above a waterfall. Over time, in both experiments the LP fish evolved phenotypes and life histories similar to those of natural LP guppies (Reznick & Bryga, 1987; Reznick *et al*., 1997; Kemp *et al*., 2009).

In the previous chapter, patterns of neutral variation and demographic histories of these experimental populations and a natural LP population were investigated. We found that ECLP was likely founded by an already bottlenecked HP population (ECHP), and subsequently experienced another bottleneck, resulting in low overall diversity and a small current effective population size (N_E). TULP on the other hand was founded from a large, stable population (GHP), and the introduced population maintained high levels of genetic variation and likely experienced rapid population growth. Finally, demographic modelling revealed the natural LP population (GLP) was likely founded by very few individuals, resulting in extremely low genetic variation and the smallest N_E of the populations under investigation.

In this chapter, we use whole genome sequencing data of these six populations to search for signals of selection along the genome, and we use the inferred demographic history of the populations to test ideas about the likelihood of molecular convergence.

3.3 Methods

Genomic data

Analyses in this chapter used the same genomic data set as described in chapter 2. Briefly, approximately 20 fish were sampled from each location in the spring of 2013. DNA was extracted using Qiagen DNeasy Blood and Tissue kit, and whole genome sequences were obtained with Illumina HiSeq technology. Raw data was processed using the Best Practices protocol as provided by GATK

(v3.8 & 4.0) (van der Auwera *et al.*, 2014). A more detailed description of this protocol can be found in chapter 2. The final VCF file contained 5,389,971 SNPs. For analyses based on haplotypes, VCF files were phased as in Malinsky *et al*. (2018), using BEAGLE (Browning and Browning, 2007) and SHAPEIT2 (Delaneau *et al*., 2012) for optimum phasing quality.

Outlier window analysis

To identify putatively locally adaptive loci between HP-LP pairs, we scanned the genome using multiple statistics, in an outlier window approach. We used the GHP/TULP, instead of the TUHP/TULP comparison because we were interested in selection since the introduction event. Each of the measures used here differ slightly in their approach to detecting signals of selection. We use allele frequency (ΔAF) between the HP and LP populations to detect shifts in allele frequencies, F_{ST} to identify strong local population differentiation, changes in nucleotide diversity ($\Delta\pi$) between HP and LP populations to find regions with a loss of genetic diversity, and finally XP-EHH as a method to detect regions with extended haplotype homozygosity. Additionally, as Whiting and Fraser (2020) found through simulated datasets that $\Delta\pi$ decreases in informativeness between populations that have diverged beyond 3000 generations, whereas D_{XY} becomes more informative after this time point, we calculated D_{XY} for GHP/GLP instead of $\Delta\pi$.

The allele frequency of the HP source population was used as the starting frequency of the introduced LP populations, allowing us to track changes in allele frequency over time in the LP populations. Here, ΔAF is calculated as:

$$|\Delta AF| = abs(AF_{LP} - AF_{HP})$$

For per SNP analysis of ΔAF in outlier windows, ΔAF was calculated as $AF_{LP} - AF_{HP}$, to obtain information on the direction of the allele frequency change in the outlier window. Likewise, comparing the ratio of π (Nei and Li, 1979) between diverging populations can reveal regions under selection where one population has reduced π compared to the other population. In this study LP populations and their HP source are considered diverging populations and Δπ is calculated as:

$$\Delta\pi = \log_{10}\left(\frac{\pi_{HP}}{\pi_{LP}}\right)$$

F_{ST} (Hudson *et al*., 1992; Weir & Cockerham, 1984) is a commonly used measure of relative differentiation and is maximised when a genomic region has lowest within population variance and highest between population variance. Finally, D_{XY} is a measure of absolute divergence. It is similar to π, except that π is calculated as the number of pairwise differences between sequence comparisons within a population, and D_{XY} is calculated between sequences of two populations (Nei, 1987). Nucleotide diversity, $F_{ST,}$ and D_{XY} were calculated in PopGenome (Pfeifer *et al*., 2014), across 75kb non-overlapping windows along the 23 chromosomes and the unplaced scaffolds. Per SNP allele frequencies were obtained with VCFTOOLS v12.b (Danecek *et al.*, 2011), and converted to non-overlapping windows of the same sizes as the other measures. This window size was chosen based on LD decay analysis with PopLDdecay (Zhang *et al.*, 2018). At 75kb, r^2 decays below 0.2 in GHP, TULP and ECHP (supplementary figure S3.1). For GLP and ECLP however, r^2 does not decay as rapidly and even at 500kb it is still above 0.2.

Windows were identified as outliers if their respective value for |ΔAF|, F_{ST}, Δπ or D_{XY} were outside of a population specific, genome-wide 95% threshold.

In addition to site frequency statistics we used a haplotype based method, XP-EHH, to detect positive selective sweeps in which the selected allele has approached or achieved fixation in one population but is still polymorphic in the population as a whole (Sabeti *et al.*, 2007). XP-EHH compares haplotype lengths across two populations to control for local variation in recombination rates, and by normalising XP-EHH values it corrects for genome-wide differences in haplotype length between populations. XP-EHH was calculated in SELSCAN v1.2.0a (Szpiech and Hernandez, 2014), using the default settings, except for a MAF filter of 0.01, and frequency-normalised over all scaffolds using the normalisation script provided by the software. Outlier windows were defined by a value of XP-EHH > 2 (following Marques *et al.*, 2018). Because GLP and TULP share the same HP source, we calculated Tajima's D for these windows using PopGenome to investigate whether shared overlapping outlier windows in GLP and TULP stem from selection events in GHP rather than convergence in the LP populations. We then calculated the median and the median absolute deviation (MAD) of Tajima's D. To accommodate for skew in the distributions of Tajima's D, MAD was calculated using a double MAD approach where MAD is calculated separately for values smaller than the median and values larger than the median. To investigate signals of selection within rivers, the number of overlapping outliers among measures was calculated using the R package SuperExactTest (Wang *et al.*, 2015). This package also calculates the expected number of overlapping windows for each set based on a hypergeometric distribution.

3.4 Results

Outlier window analysis

We used an outlier window approach with |ΔAF|, F_{ST}, $\Delta\pi$ (D_{XY} for GLP instead of $\Delta\pi$), to identify signals of selection across the genome between experimental populations and their sources and one natural HP/LP pair for comparison. A total of 9227 windows of 75kb length on 23 chromosomes, and 667 windows on unplaced scaffolds were processed with PopGenome (Pfeifer *et al.*, 2014) and VCFTOOLS v16 (Danecek *et al.*, 2011). We found that mean F_{ST} and |ΔAF| were highest in GHP/GLP (genome wide F_{ST}=0.307 and |ΔAF|=0.134, table 3.1, figure 3.1). Lowest mean |ΔAF| was found in ECHP/ECLP (|ΔAF|=0.058, table 3.1, figure 3.1), however lowest mean F_{ST} was found in GHP/TULP (F_{ST}=0.061). The distribution of F_{ST} in ECHP/ECLP (figure 3.1), with it bordering on 0 and then a long tail towards larger values could be explained by the young age of a bottlenecked population but with migration occurring between ECHP and ECLP (see Chapter 2). Genome-wide $\Delta\pi$ was positive in ECHP/ECLP, indicating that ECLP had lower values of π compared to ECHP. In the cross-drainage introduction GHP/TULP however, we saw the opposite signal, with a slightly negative value for $\Delta\pi$, indicating π had increased in TULP since its introduction from GHP (see Chapter 2).

Outlier windows for |ΔAF|, F_{ST}, $\Delta\pi$, and D_{XY} were identified using a genome-wide 95% confidence interval (table 3.1) and therefore each population had the same number of outlying windows per statistic. We used a cut-off of >2 for the normalised XP-EHH values and can therefore compare number of outlier windows in each population pair as a measure of positive selection. We found that the number of outlier XP-EHH windows was highest in ECHP/ECLP (N=288) followed by GHP/TULP (N=118) and finally GHP/GLP (N=91).

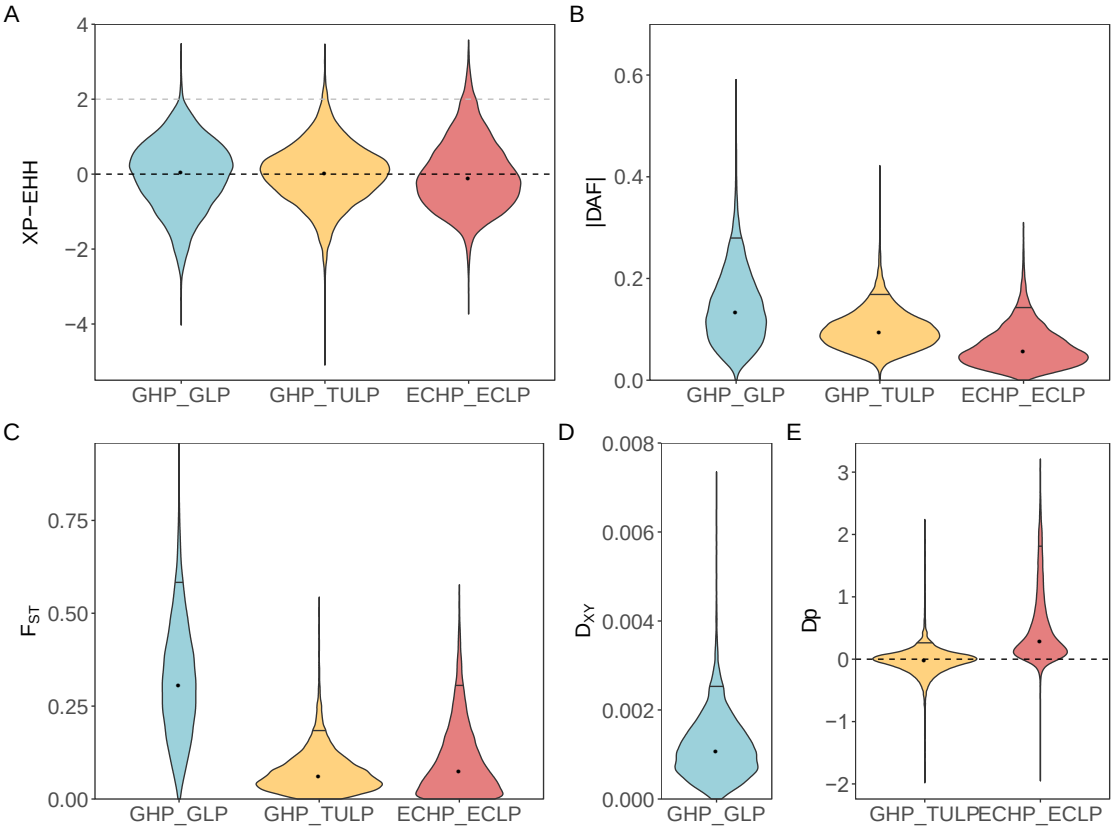


Figure 3.1 Distributions of genome scan measures for A) XP-EHH, B) |ΔAF|, C) F_{ST}, D) D_{XY} (GLP) and E) $\Delta\pi$, (TULP and ECLP)

Table 3.1 Summary of genome scan results for ΔAF, F_{ST}, Δπ and D_{XY} in 75kb windows.

Population	mean |ΔAF|	median |ΔAF|	95% CI	mean F_{ST}	median F_{ST}	95% CI	mean Δπ	median Δπ	95% CI	mean D_{XY}	median D_{XY}	95% CI
GHP/GLP	0.145	0.134	0.278	0.316	0.307	0.582	NA	NA	NA	0.001	0.001	0.003
GHP/TULP	0.099	0.095	0.169	0.074	0.061	0.184	-0.026	-0.014	0.265	NA	NA	NA
ECHP/ECLP	0.066	0.058	0.142	0.103	0.075	0.300	0.508	0.293	1.813	NA	NA	NA

Evidence of selection within rivers

To explore signals of positive selection within a river, we investigated overlapping outlier windows among measures within each population pair, compared to what is expected by random chance. In all population pairs, comparisons including all four measurements had more overlapping windows than might be expected by chance (highest fold enrichment, figures 3.2-3.4). We found that ECHP/ECLP had the highest fold enrichment, regardless of which statistics were compared, and had the most overlapping outlier windows for each comparison, except for $|\Delta AF|$ and F_{ST}, for which GHP/TULP had more overlap. GHP/GLP had few overlapping windows in comparisons that considered XP-EHH.

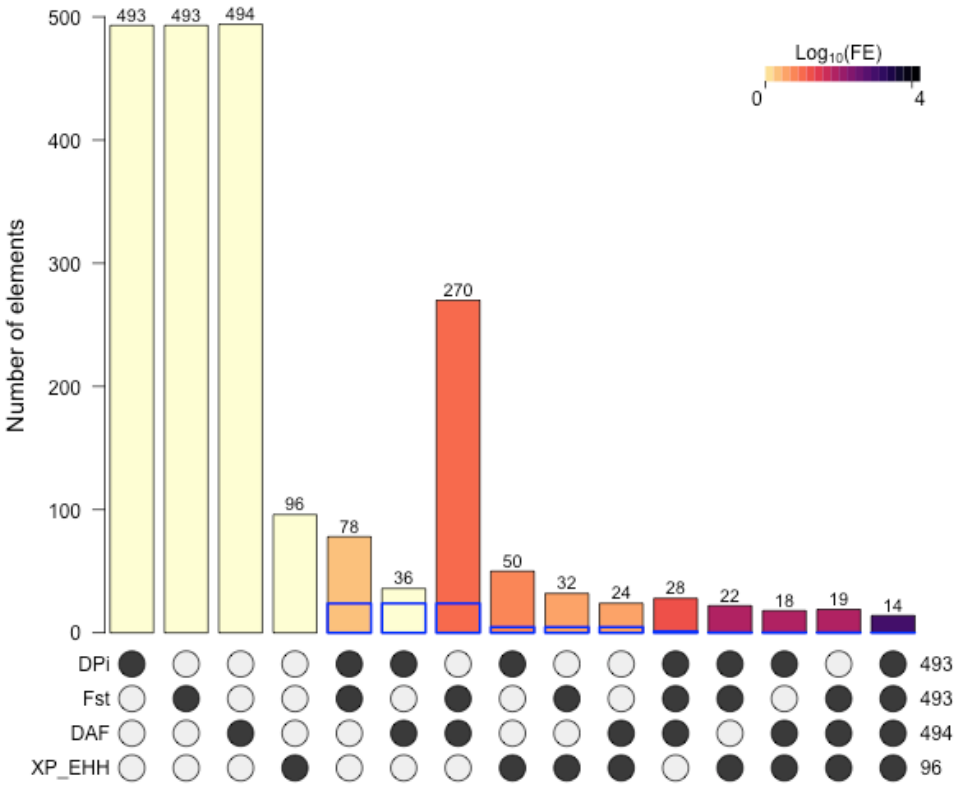


Figure 3.2 Number of overlapping outlier windows among all comparisons in TULP/GHP. Numbers on top of the bars are the overlapping windows, blue boxes indicate the expected number of overlapping windows for that comparison. Colour scale indicates log-scaled fold enrichment.

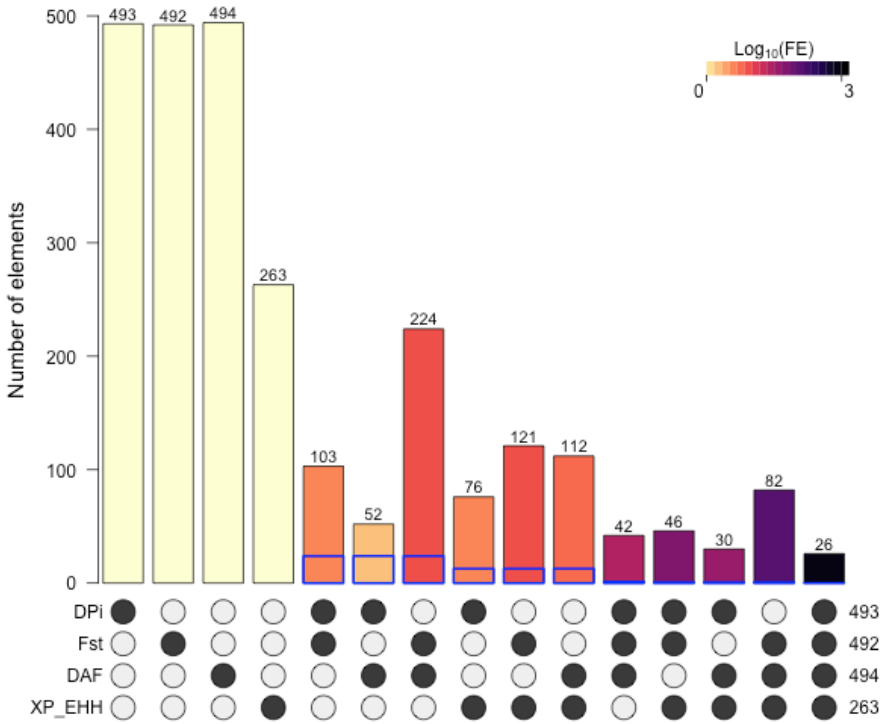


Figure 3.4 Number of overlapping outlier windows among all comparisons in ECLP/ECHP. Numbers on top of the bars are the overlapping windows, blue boxes indicate the expected number of overlapping windows for that comparison. Colour scale indicates log-scaled fold

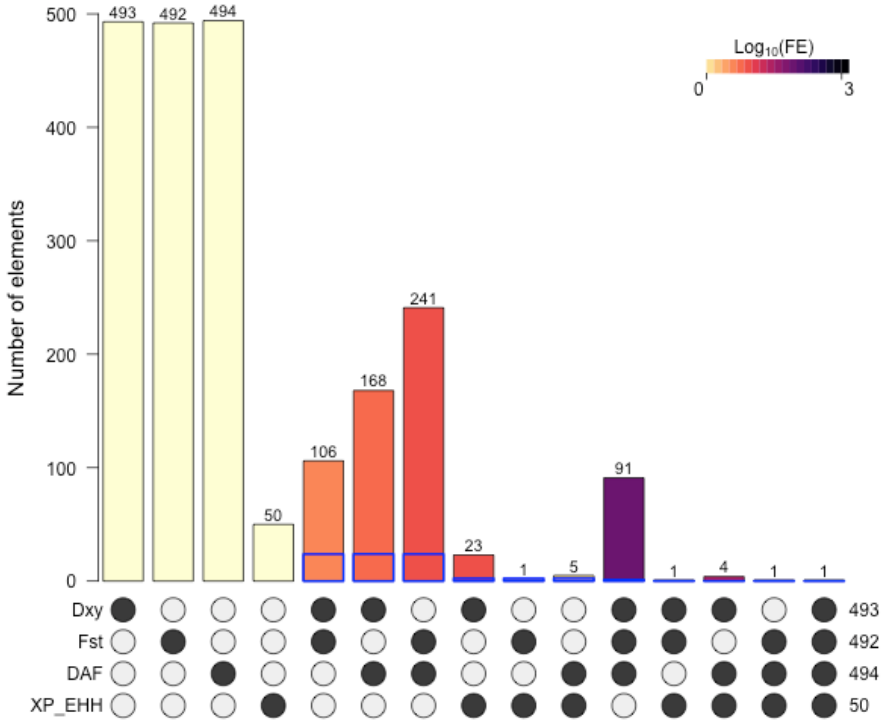


Figure 3.3 Number of overlapping outlier windows among all comparisons in GLP/GHP. Numbers on top of the bars are the overlapping windows, blue boxes indicate the expected number of overlapping windows for that comparison. Colour scale indicates log-scaled fold

By intersecting the outlier lists of the non-haplotype-based scans (|ΔAF|, F_{ST} and Δπ) and comparing the values of XP-EHH of these windows, we can identify if the detected signals reflect recent sweeps or more ancient sweeps. We found that in GHP/TULP and ECHP/ECLP the overlapping outlier windows also had increased haplotype homozygosity compared to a random sample of XP-EHH values (p-value < 0.0001, Mann-Whitney U test, figure 3.5b&c), suggesting recent positive selective sweeps occurred in these populations. When we intersected overlapping outlier windows of D_{XY}, |ΔAF| and F_{ST} for GHP/GLP, and compared them to XP-EHH values for those windows, we found GHP/GLP did not have significantly higher values of XP-EHH than a random sample (p-value = 0.1046, Mann-Whitney U test, figure 3.5a).

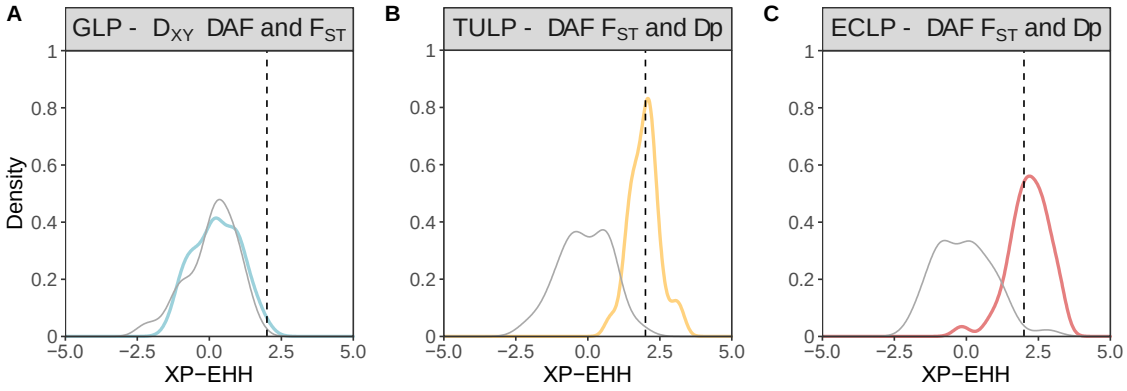


Figure 3.5 Distribution of XP-EHH for overlapping outlier positions (coloured line) compared to a random sample of equal size (grey line). A) GLP/GHP, overlapping outliers of D_{XY}, |ΔAF| and F_{ST} were compared to XP-EHH, B) TULP/GHP C) ECLP/ECHP outlier lists of |ΔAF|, F_{ST} and Δπ were compared to XP-EHH. Vertical dashed line is the outlier cut-off of XP-EHH.

Intersecting lists of all four measurements to identify signals of selection within a river resulted in very few windows, and may be overly conservative. Instead, we define candidate regions under selection if it is an outlier in three of the four measures we used. In GHP/TULP this resulted in 45 overlapping outlier windows using the 95% cut-off, spread across ten chromosomes and three unplaced scaffolds (supplementary table S3.1). Nearly a quarter of these windows (N=10) were found on chromosome 1 (figure 3.6), with three consecutive windows from 14700001 to 14925000 bp. Examining this chromosome in the other pairs revealed that GHP/GLP had five overlapping windows among $|\Delta AF|$, F_{ST} and D_{XY} on chromosome 1, but none of these directly overlapped with the outlier windows in GHP/TULP or ECHP/ECLP. Only one outlier window on chromosome 1 overlapped among three measures in ECHP/ECLP and this window did not appear in the overlapping outliers in the other two comparisons. Using the 99% cut-off, we found 1 window in GHP/TULP on chromosome 1 (14700001-14775000 bp) and no outlier windows in GHP/GLP and ECHP/ECLP. We calculated per SNP values of all measures within these three windows to further investigate candidate genes under selection in GHP/TULP. This window contains fourteen genes, with high differentiation across the entire region (figure 3.7).

The next most consistent outlier region in GHP/TULP was nine windows on unplaced scaffold 94 and two of the windows on chromosome 20 (supplementary table S3.1). Whiting *et al*. (2020) placed this scaffold on chromosome 20 using HiC mapping and describe two haplotypes stretching across the first 3Mb of chromosome 20 (with scaffold 94 placed) that are polymorphic in the Caroni drainage (where Guanapo and El Cedro are located) and show signatures of selection in LP populations. The outliers on chromosome 20 and scaffold 94

reported here fall in adjacent regions when we place the scaffold on chromosome 20 (figure 3.8), supporting the hypothesis of a large haplotype under selection in this region in at least GHP/TULP and GHP/GLP. There were no overlapping outliers at the 99% quantile between three of the four measures for any of the population pairs.

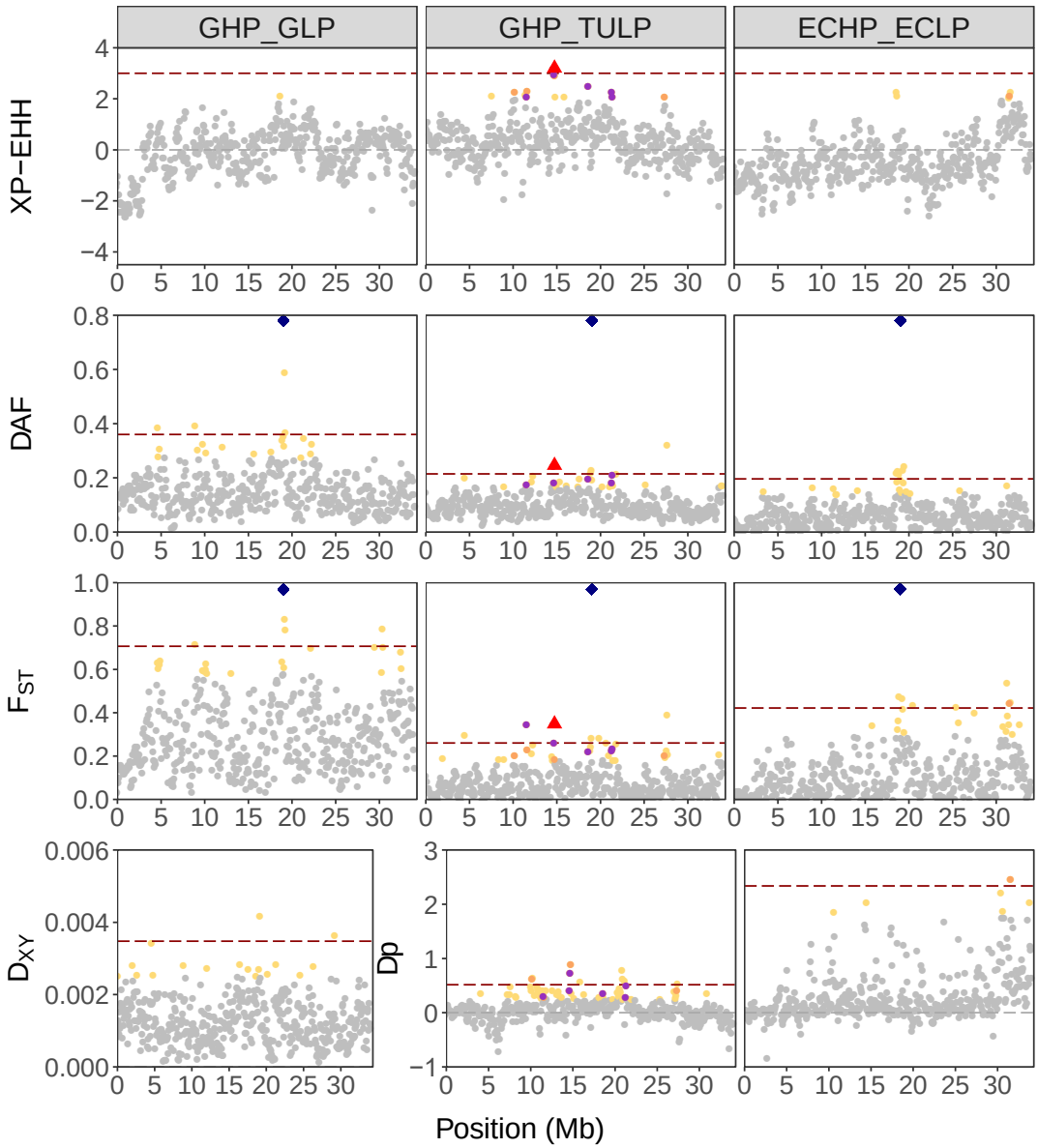


Figure 3.6 Overlapping outlier windows on chromosome 1 in all three populations. Circles indicate 95% quantile outliers, triangles indicate 99% quantile outliers. Yellow is unique outliers per population, orange is overlapping outlier windows among three of the four measures, purple marks overlapping outlier windows among all measures at the 95% quantile. Red triangle marks the overlapping window for 3 of the 4 measures at the 99% quantile. Dark red dashed line indicates the 99% quantile cut-off. Blue diamonds at the top of the |ΔAF| and F_{ST} plots indicates the window that overlaps among all populations at the 95% quantile.

Figure 3.7 Per SNP values for the three consecutive outlier windows on chromosome 1 (14700001-14925000 bp) with 250kb windows on either side for TULP/GHP. Values for $\Delta\pi$ are given in 1kb windows. Outlier windows are placed between the vertical dashed lines. Yellow points mark 95% confidence interval outliers.

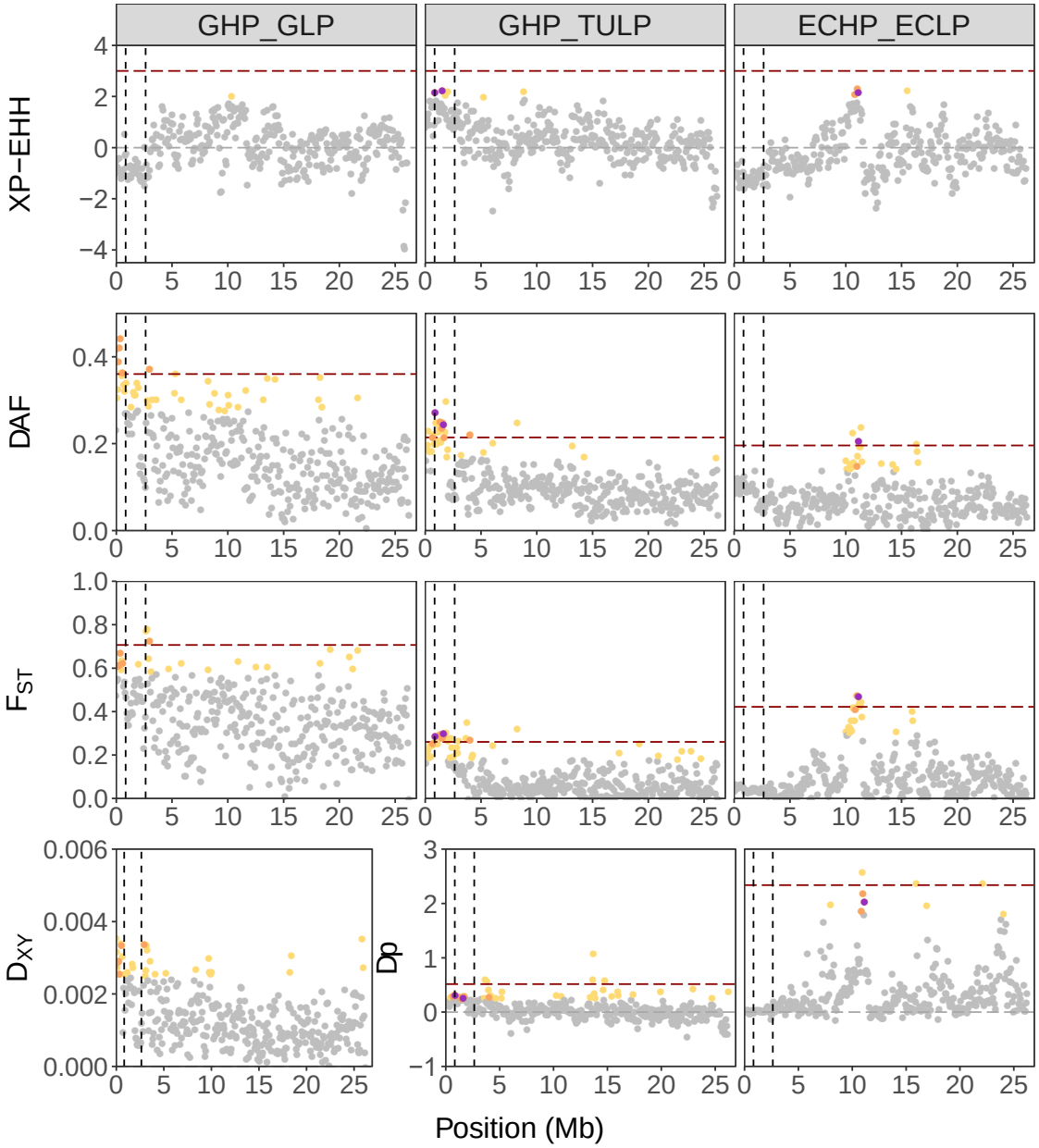


Figure 3.8 Overlapping outlier windows on chromosome 20/scaffold 94 in all three populations. Circles indicate 95% quantile outliers, triangles indicate 99% quantile outliers. Yellow is unique outliers per population, orange is overlapping outlier windows among three of the four measures, purple marks overlapping outlier windows among all measures at the 95% quantile. Dark red dashed line indicates the 99% quantile cut-off.

In ECHP/ECLP, we found 118 overlapping outlier windows, distributed across sixteen chromosomes and two unplaced scaffolds (supplementary table S3.2). Chromosome 8 had the most overlapping windows (N=28), followed by chromosome 10 and 12 (N=19 for both). On chromosome 8 we found a cluster of five consecutive overlapping windows from 21075000-21450000 bp in

ECHP/ECLP (figure 3.9). This region contained 16 genes, with per SNP outlier values evenly spread across them (figure 3.10). Interestingly, one of the genes in this region is melanoregulin (MREG), which is related to melanosome distribution in cells (Wu *et al.*, 2012). There were no overlapping outliers in any of the population pairs when using the 99% cut-off.

We noticed an interesting pattern in $|\Delta AF|$ in this region, where we observe a block with few values between 0 and 0.5. Investigating further by calculating heterozygosity for males and females, we found this is caused by a lack of heterozygote individuals in ECLP across large parts of the chromosome, but especially between 15 Mb to 21.5 Mb in males, suggesting most SNPs have either been fixed or lost in ECLP since the introduction (figure 3.11). We found both ECHP and ECLP males have genome-wide reduced heterozygosity compared to females (table 3.2, P=0.0001 and p= 0.031 respectively, Mann-Whitney U test). In outlier regions specifically, ECLP observed heterozygosity was reduced compared to ECHP (p=0.044), and the difference between ECHP and ECLP was greater in males than in females (p<0.001, Mann-Whitney U test).

Table 3.2 Observed heterozygosity in the outlier regions and a random sample of ECLP and ECHP. The genome wide value is based on the average of 500 random samples the same size as the number of outlier regions.

		Observed Heterozygosity	
Population	**Sex**	**random sample**	**outlier regions**
ECLP	Males	0.0719	0.0099
	Females	0.0728	0.0189
ECHP	Males	0.1143	0.2090
	Females	0.1339	0.1916

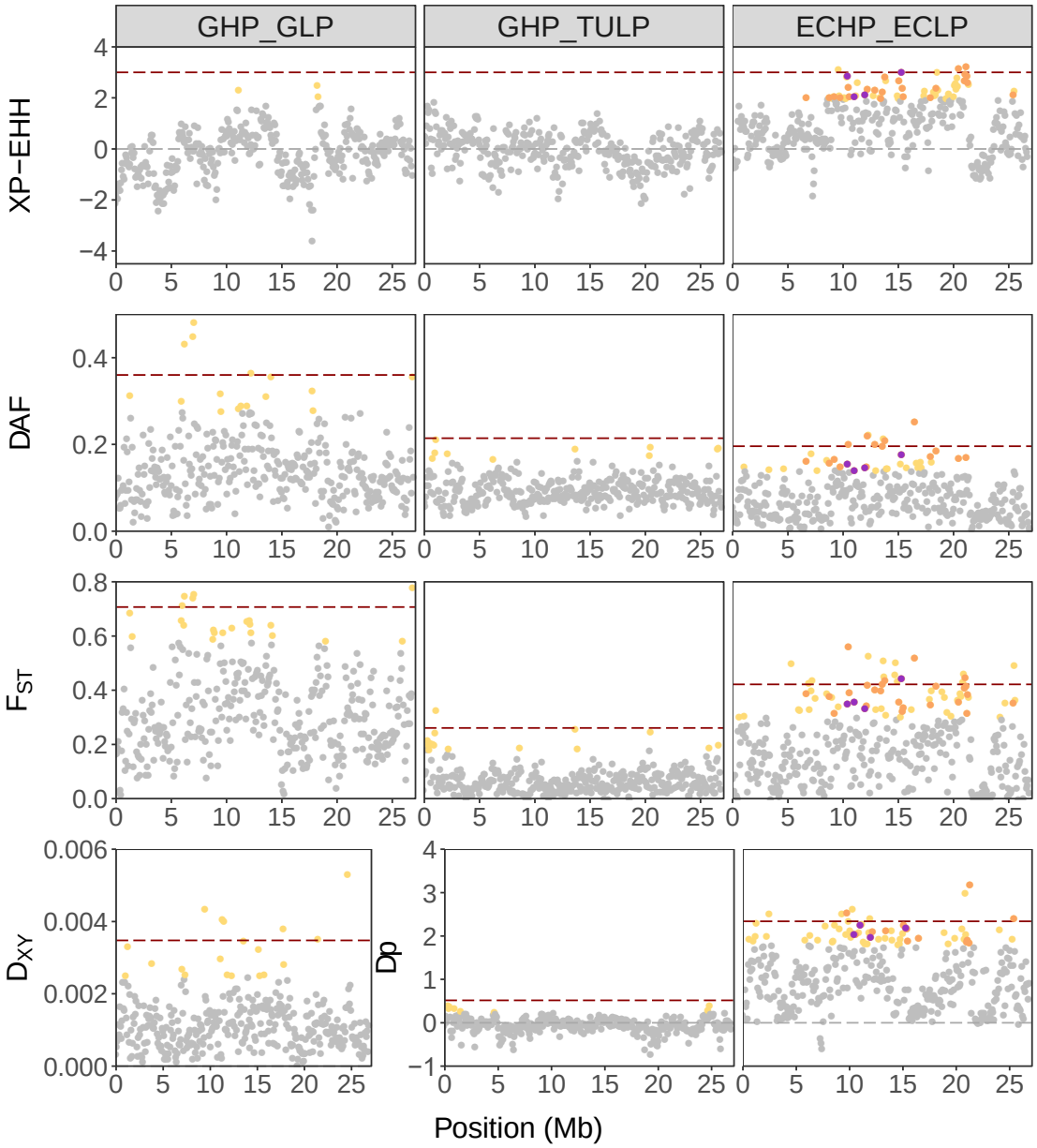


Figure 3.9 Overlapping outlier windows on chromosome 8 in all three populations. Yellow points mark 95% confidence interval outliers, orange points overlapping outlier windows within a population among three of the four measures used, and purple points mark overlapping outlier windows among all four measures. Dark red dashed line indicates the 99% quantile cut-off.

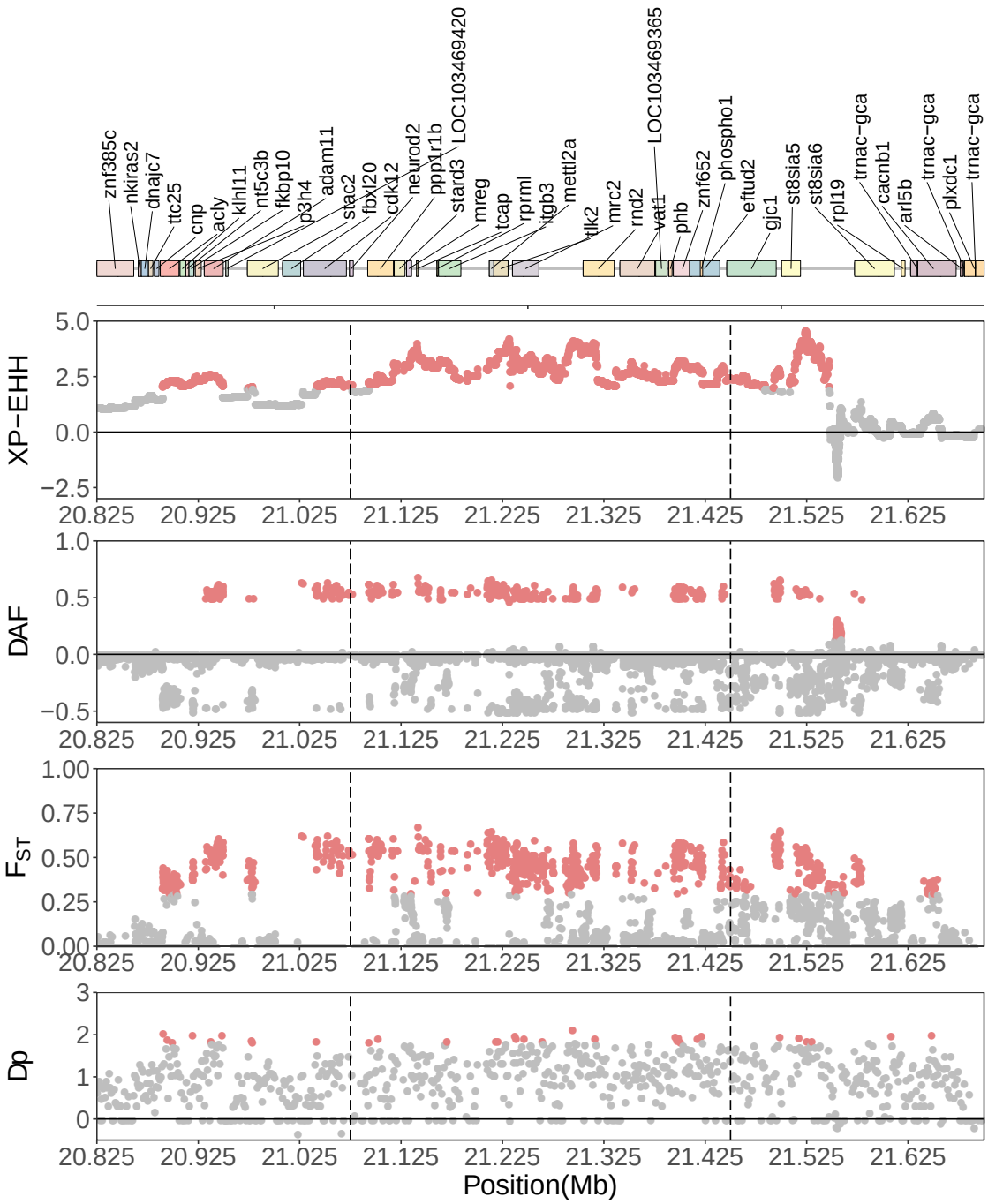


Figure 3.10 Per SNP values for the three consecutive outlier windows on chromosome 8 (21075000-21450000 bp) with 250kb windows on either side for ECHP/ECLP. Values for $\Delta\pi$ are given in 1kb windows. The outlier windows are placed between the vertical dashed lines. Red points mark 95% confidence interval outliers.

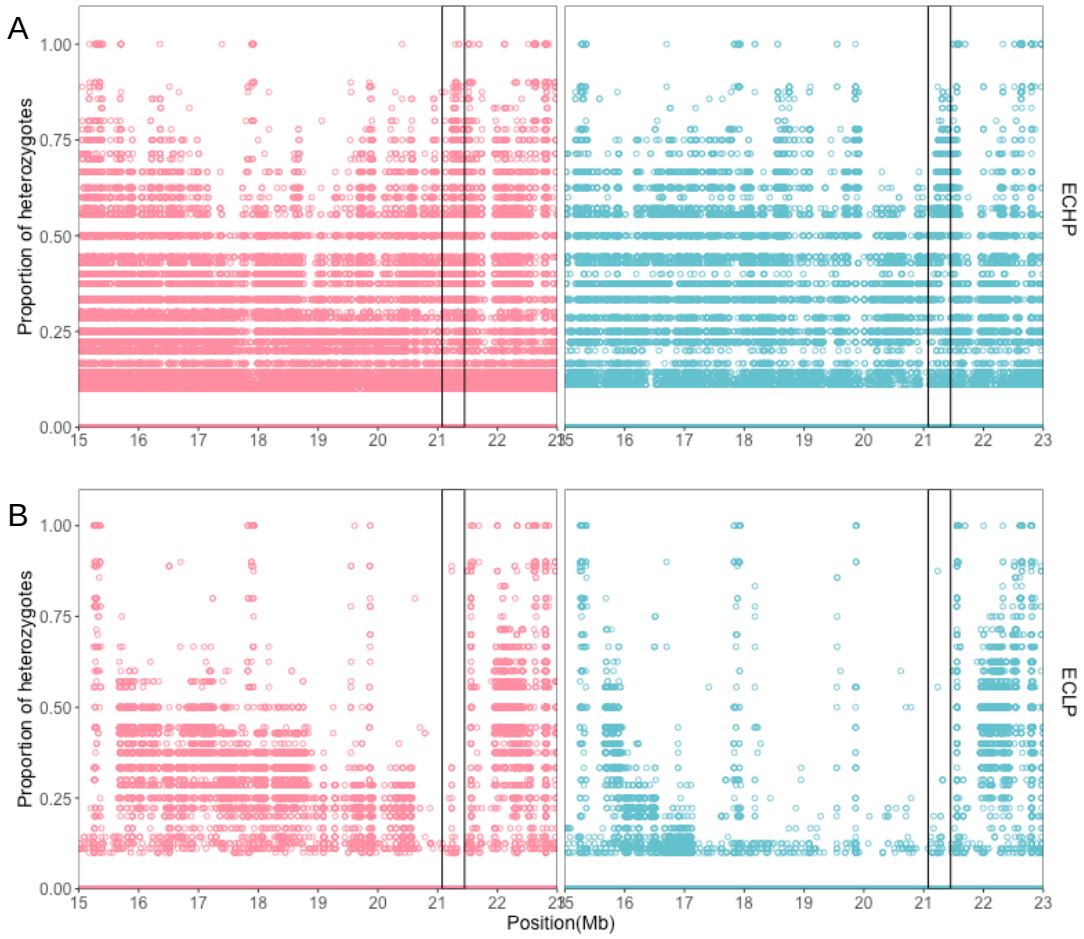


Figure 3.11 Proportion of heterozygotes for a region on chromosome 8, females (pink) and males (blue). Values are shown for A) ECHP and B) ECLP. Black rectangle shows the position of the three consecutive overlapping outlier windows plotted in figure 10 (21075000-21450000 bp).

The next notable region of selection in the ECLP/ECHP pair is on chromosome 12, where we find a cluster of six consecutive outlier windows from 9000001 to 9450000 bp at the 95% confidence interval (figure 3.12). One window overlapped for three of the four measures at the 99% confidence interval for ECLP/ECHP (9000001-9075000 bp) and no windows overlapped among all four measures. When we calculated per SNP values for these windows we saw that outliers were relatively equally distributed across the region (figure 3.13). We also noticed a similar pattern in $|\Delta AF|$ as we saw for chromosome 8 and plotting observed heterozygosity again showed an extreme loss of heterozygotes, with strong breakpoints in the males extending from 8.7 Mb to 11.5 Mb (figure 3.14). Although

differentiation was high across the region, there are several interesting candidate genes in this window, including PTGES (prostaglandin synthase) and PTGER4 (a prostaglandin receptor) and ubiquitin carboxyl-terminal hydrolase 20 (USP20).

Because of the high level of homozygosity in GLP, we excluded XP-EHH from our overlapping outlier analysis, instead we identify a window as a selection candidate if it is an outlier in D_{XY}, $|\Delta AF|$ and $F_{ST.}$ This resulted in 91 overlapping outlier windows on twenty-one chromosomes and one unplaced scaffold (supplementary table S3.3). The highest density of outliers was found on chromosome 20, with a total of 10 overlapping windows of which 6 were located in the first Mb of the chromosome (supplementary table S3.3, figure 3.8).

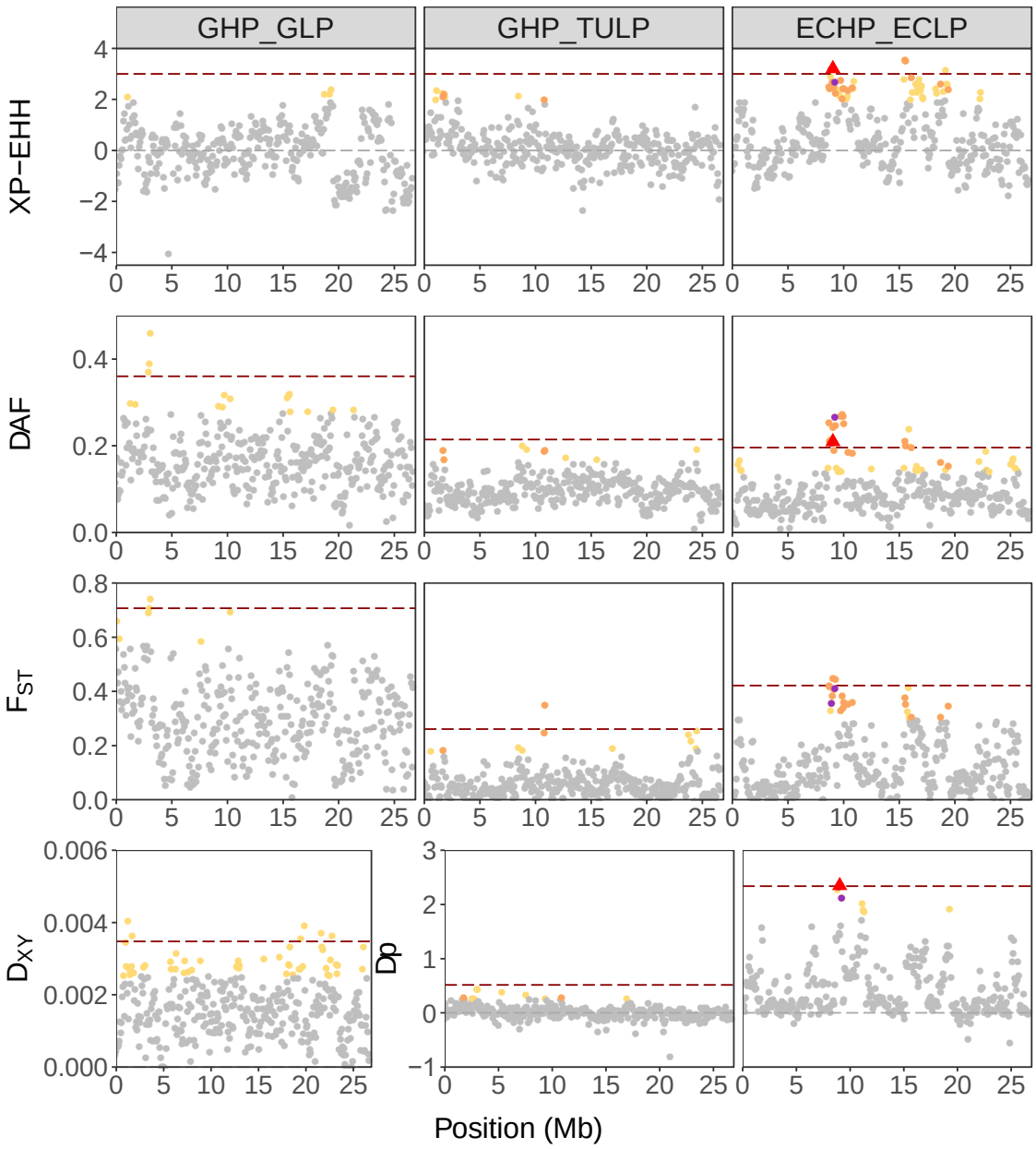


Figure 3.12 Overlapping outlier windows on chromosome 12 in all three populations. Circles indicate 95% quantile outliers, triangles indicate 99% quantile outliers. Yellow is unique outliers per population, orange is overlapping outlier windows among three of the four measures, purple marks overlapping outlier windows among all measures at the 95% quantile. Red triangle marks the overlapping window for 3 of the 4 measures at the 99% quantile. Dark red dashed line indicates the 99% quantile cut-off

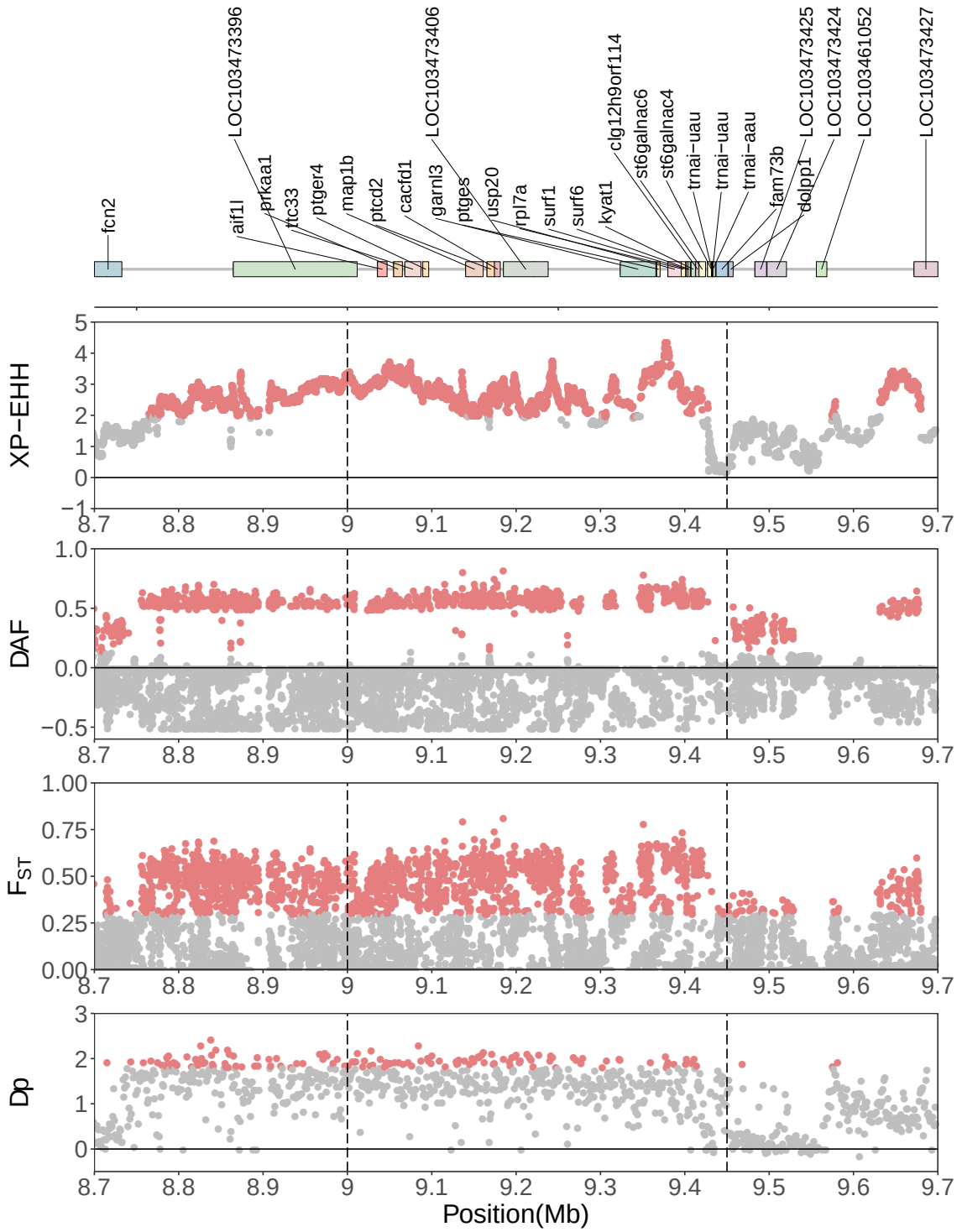


Figure 3.13 Per SNP values for the four consecutive outlier windows on chromosome 12 (9000001 to 9450000 bp) with 250kb windows on either side for ECHP/ECLP. values for $\Delta\pi$ are given in 1kb windows. Outlier windows are located between the vertical dashed lines. Red points mark 95% confidence interval outliers.

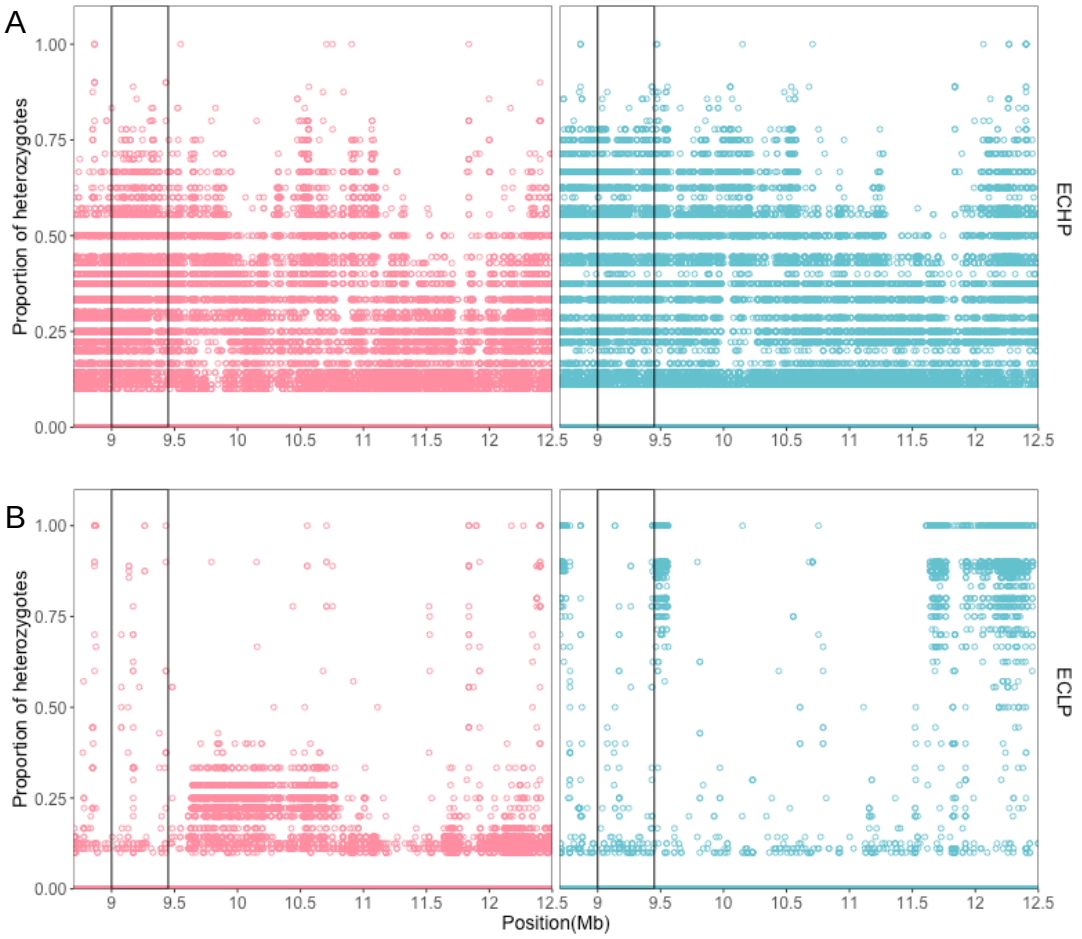


Figure 3.14 Proportion of heterozygotes for a region on chromosome 12, females (pink) and males (blue. Values are shown for A) ECHP and B) ECLP. Black rectangle shows the position of the four consecutive overlapping outlier windows plotted in figure 9 (9000001 to 9450000 bp).

Convergent evolution between rivers

We next looked for signals of convergent evolution between river pairs by intersecting outlier window lists of each measure for all comparisons. Most outliers were unique to a river (table 3.3). When considering overlapping outlier windows in pairwise comparisons, GHP/GLP and ECHP/ECLP had the most overlapping windows per measure, and overlapped more than expected by chance alone for |ΔAF| (p-value < 0.0001, SuperExactTest in R), F_{ST} (p-value=0.0008), and XP-EHH (p-value = 0.0003, table 3.3). GHP/GLP and GHP/TULP overlapped more than expected for |ΔAF| (p-value < 0.0001) and F_{ST}

(p-value=0.0138), but not for XP-EHH (p-value = 0.0810). Finally, ECHP/ECLP and GHP/TULP had the least number of overlapping outliers for |ΔAF| and F_{ST}, although, this was still more than expected for |ΔAF| (p-value < 0.0001), but not for F_{ST} (p-value = 0.2029, table 3.3). As we calculated D_{XY} instead of $\Delta\pi$ for GLP, only the comparison between ECLP and TULP could be made for $\Delta\pi$, which resulted in 15 overlapping windows, which was not more than expected by chance (p-value = 0.99, table 3.3).

Table 3.3 Number of observed outliers compared among populations. Where GLP is involved, no comparisons with Δπ could be made. Expected overlap in brackets and significant overlap printed in bold.

	Comparison	|ΔAF|	F_{ST}	XP-EHH	Δπ
Unique	GLP	364	417	40	NA
	ECLP	383	424	254	493
	TULP	397	430	91	493
Pairwise	GLP-ECLP	**78 (24.71)**	**41 (24.57)**	**7 (1.31)**	NA
	GLP-TULP	**64 (24.71)**	**36 (24.63)**	2 (0.47)	NA
	ECLP-TULP	**45 (24.71)**	29 (24.63)	2 (2.54)	15 (24.68)
All	GLP-ECLP-TULP	**12 (1.24)**	2 (1.23)	0 (0.01)	NA

When comparing all three rivers, twelve |ΔAF| overlapping outliers windows were found, which was more than expected to overlap by chance alone (p-value < 0.0001) and were distributed across six chromosomes and two unplaced scaffolds (table 3.4). Two F_{ST} outlier windows overlapped, one on chromosome 1 (18975001-19050000) (figure 3.6) and one on chromosome 13 (17850001-17925000), however this was not significantly more than expected by chance (p-value = 0.3486). The window on chromosome 1 was also found to be overlapping among all populations for |ΔAF| (figure 3.6). No windows overlapped among all three populations for XP-EHH.

Table 3.4 Distribution of overlapping outlier windows across the genome for ΔAF among populations.

Chromosome	# of outliers
chr1	2
chr2	2
chr12	2
chr13	1
chr17	1
chr20	2
000239F_0	1
000280F_0	1

Next, we compared the outlier lists from all four measures of the introduced populations and their intersecting lists with the D_{XY} outliers from GHP/GLP making for 15 comparisons in total. We chose to use D_{XY} outliers in GHP/GLP alone as Whiting and Fraser (2020) showed it maintains a good correlation with selection regardless of population size and over longer divergent times, whereas for example, F_{ST} has a reduced correlation with selection in small populations such as GLP. Overlap among GHP/GLP outliers of D_{XY} and outliers of GHP/TULP measures was higher than expected in fourteen of the fifteen comparison (exception was D_{XY} & $\Delta\pi$ overlap), whereas overlap of GHP/GLP D_{XY} outliers with ECHP/ECLP outliers were higher than expected in ten of the fifteen comparisons (figure 3.15).

To identify candidate windows under convergent selection, we intersected the overlapping outlier list obtained in the previous section (see supplementary tables S3.1 and S3.2) with GHP/GLP D_{XY} outliers. This resulted in twelve overlapping outlier windows between GHP/GLP and GHP/TULP, exact positions of the windows on each chromosome/scaffold and the genes within the window can be found in table 3.5. Half of these windows were placed on chromosome 20/scaffold

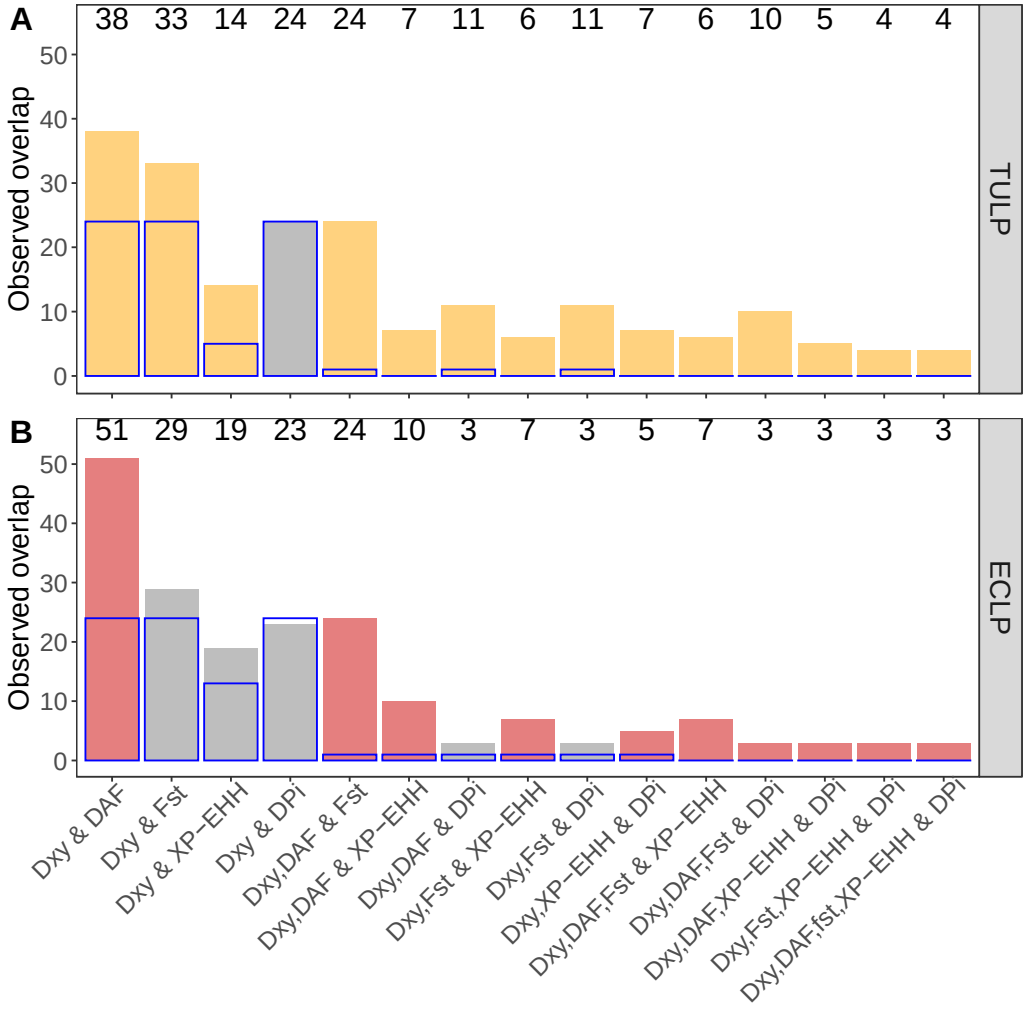


Figure 3.15 Overlapping outlier windows among outlier windows of GLP/GHP D_{XY} and from all measures of the introduced populations: A) TULP/GHP and B) ECHP/ECLP. Blue boxes indicate the expected number of overlapping windows, grey bars indicate cases where the number of observed outliers was not significantly different from the expected number.

94, which we described in the previous section as a candidate region for adaptation in TULP (figure 3.8) and was a region of interest in Whiting *et al.* (2020) investigating convergent evolution in natural populations of guppies. Another window of interest found with this analysis is chromosome 15 at roughly 5Mb, as it was also found to be an outlier in recently established introduction populations (see Chapter 5), and is within a region of interest for convergent evolution in natural populations (Fraser *et al.*, 2015; Whiting *et al.*, 2020). Calculating per SNP values for ΔAF and F_{ST}, and 1kb window values for $\Delta\pi$ and D_{XY} reveals that the majority of the SNPs in the outlier window were located in

the region that corresponds to the B-cadherin gene and CDH1 (or E-cadherin) in the adjacent window (figure 3.16). Genes in the cadherin family mediate cell-cell adhesion and have also been found regulating the migration of pigment cells (Fukuzawa & Obika, 1995; Nishimura *et al*., 1999).

Table 3.5 Positions of overlapping outlier windows among three of the four measures in either GHP/TULP or ECHP/ECLP and DXY in GHP/GLP.

Population	Chromosome	overlapping windows	Genes found
GHP/TULP-GHP/GLP	chr1	21450001-21525000	ANK3, CD027, PDE5A, PPP2R2B, WFS1, CRMP1, PCDH7
	chr6	18075001-18150000	UNC45A, RCCD1, CIB1, GDPGP1, RHCG, SLS
	chr10	4575001-4650000	IL1RAPL1, B3GALT2
	chr12	1800001-1875000	CRB2, FBXW2, WDR45, NAS-4, ZCCHC9, XRCC4
	chr15	4950001-5025000	CDH1, B-cadherin
	chr20	675001-750000	C1QL3, DTX3L
		750001-825000	RSU1, TRDMT1
	scf94	750001-825000	SNX7
		975000-1050000	LOC103482394, LOC108165682
		1275001-1350000	N/A
		1575001-1650000	N/A
	scf113	675001-750000	PFFIA1
ECHP/ECLP-GHP/GLP	chr8	13650001-13725000	TBL3, RNF151, NDUFB10, RPL3L, TEX2
		15375001-15450000	ADAP1, RAB26
	chr12	9000001-9075000	AIF1L, LAMC3, PRKAA1, TTC33
		15675001-15750000	GGT1, GGT5, CABP1, GPAT4, RAB11FIP1
		19500001-19575000	PLPPR1
	chr14	22575001-22650000	GALNT10
	chr18	3825001-3900000	SORCS2, MRC2, VSIG1, LOC103480234, H1-1, H2B-I

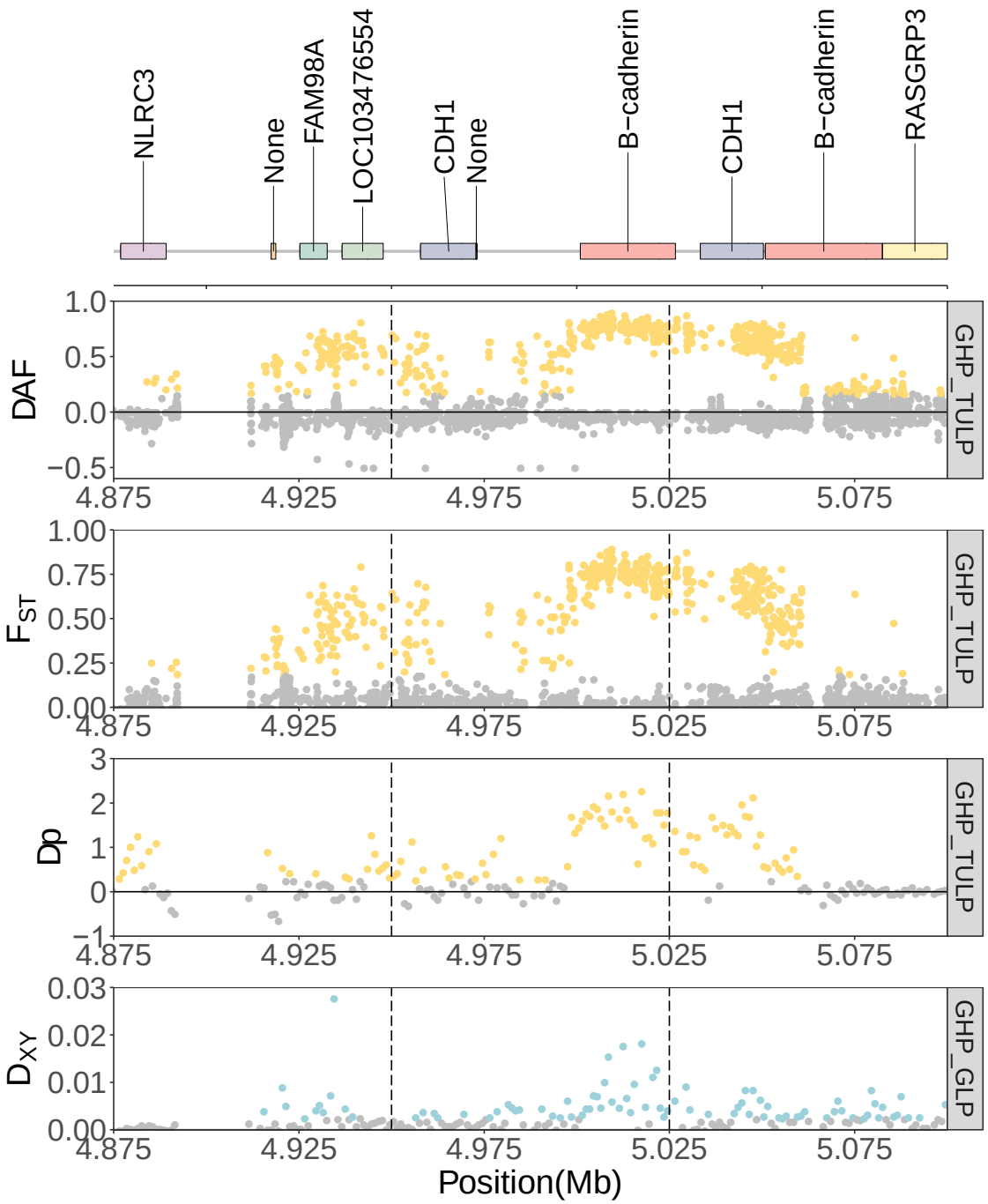


Figure 3.16 Per SNP values for the overlapping outlier window on chromosome 15 (4950000-5025000 bp), with 75kb windows on either side. Values for $\Delta\pi$ and D_{XY} are given in 1kb windows. Top three panels are ΔAF, F_{ST} and $\Delta\pi$ in TULP/GHP and bottom panel is D_{XY} in GLP/GHP. Outlier window is located between the vertical dashed lines. Yellow points mark 95% confidence interval outliers for TULP/GHP, blue points mark 95% confidence interval for GLP/GHP. Locations of the genes are plotted at the top.

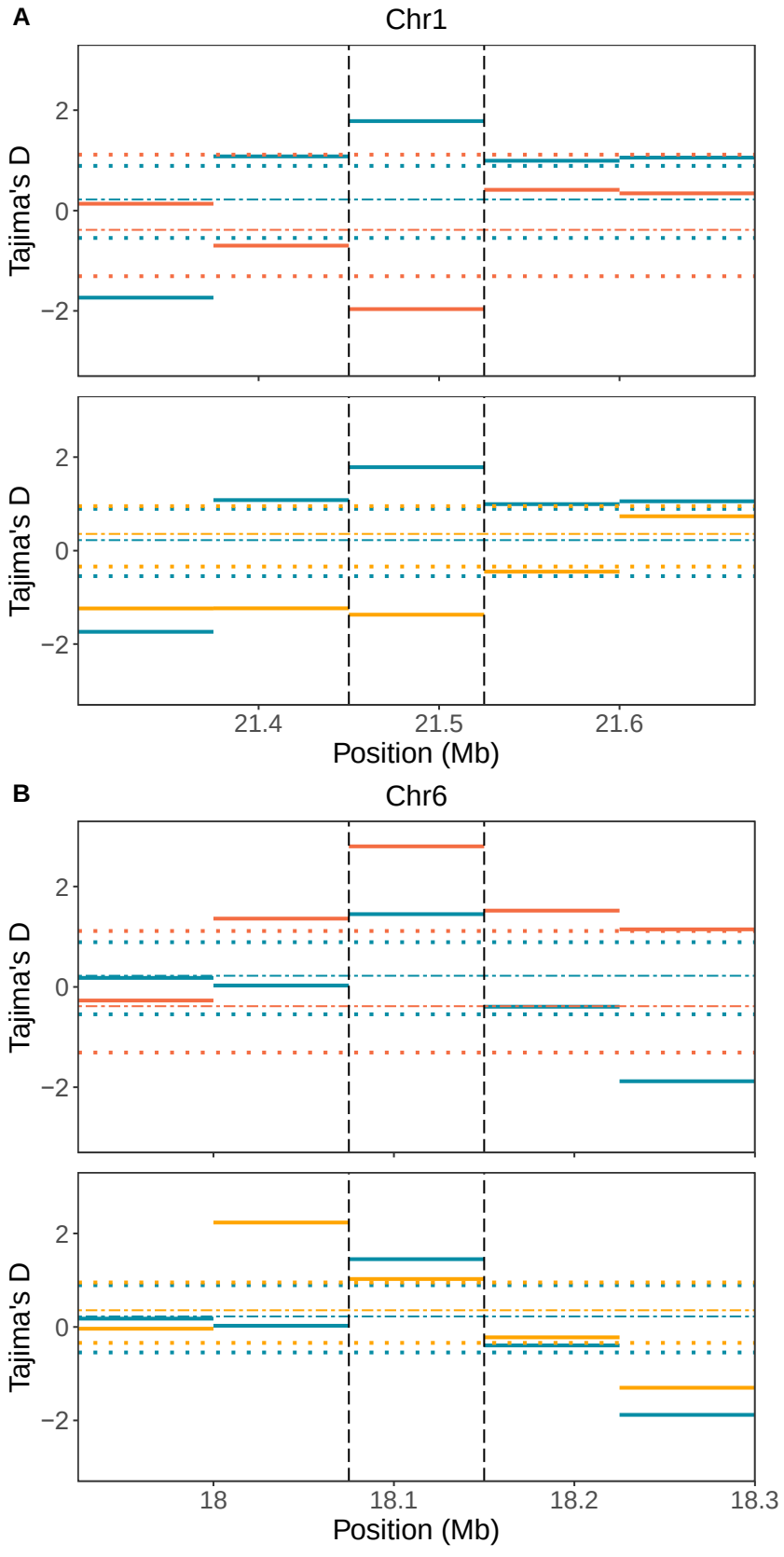
Because GLP and TULP share the same HP source, it is possible the incidence of overlap between GHP/GLP and GHP/TULP is caused by selective events in GHP driving the outliers rather than selection in GLP or TULP. To

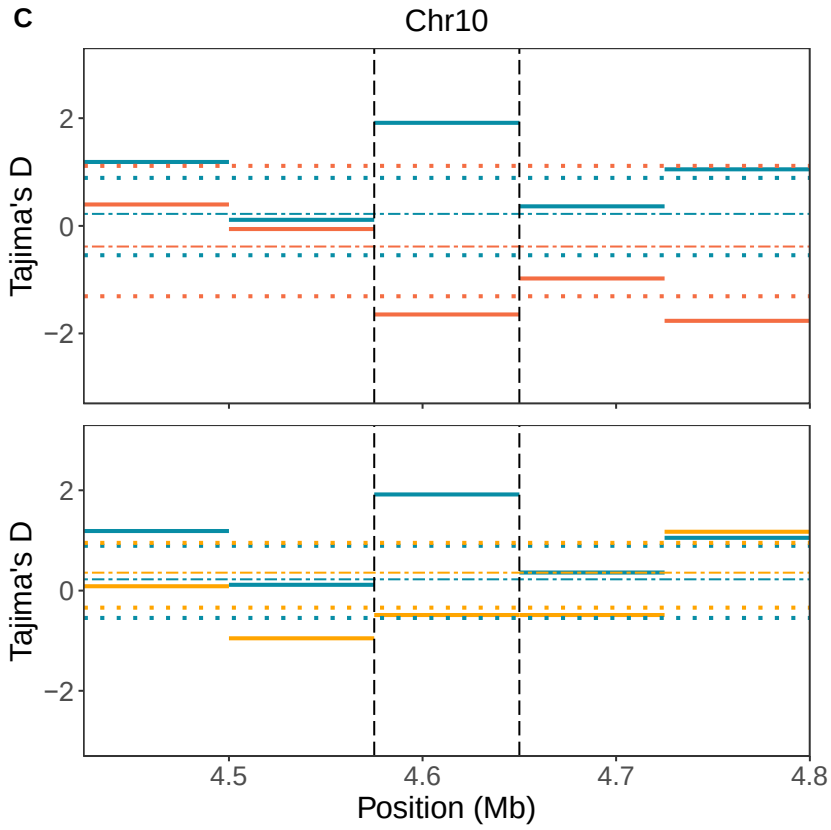
investigate this further, we calculated Tajima's D (Tajima, 1989a) for the overlapping windows to assess if selection occurred in GHP, in the LP populations, or in both (table 3.6). Because of the high level of homozygosity and the strong skew in Tajima's D in GLP it is difficult to interpret Tajima's D confidently, therefore we only focus on TULP. Based on comparing the window of interest's Tajima's D to the median absolute deviation (MAD), we propose that selection occurred in GHP only in eight of the twelve windows (figure 3.17B-D, F [750001-825000], G & H), in TULP only in one window (figure 3.17E) and in both populations in three windows (table 3.6, figure 3.17A, F [675001-750000] & H). In all candidate windows where GHP Tajima's D is different from the MAD, Tajima's D is positive, suggesting balancing selection rather than positive selection. However, tests specifically aimed at detecting balancing selection would be needed in order to draw conclusions about this. In windows where TULP Tajima's D is different from the MAD, Tajima's D is negative for all outliers except chromosome 15, suggesting positive selection (figure 3.17E).

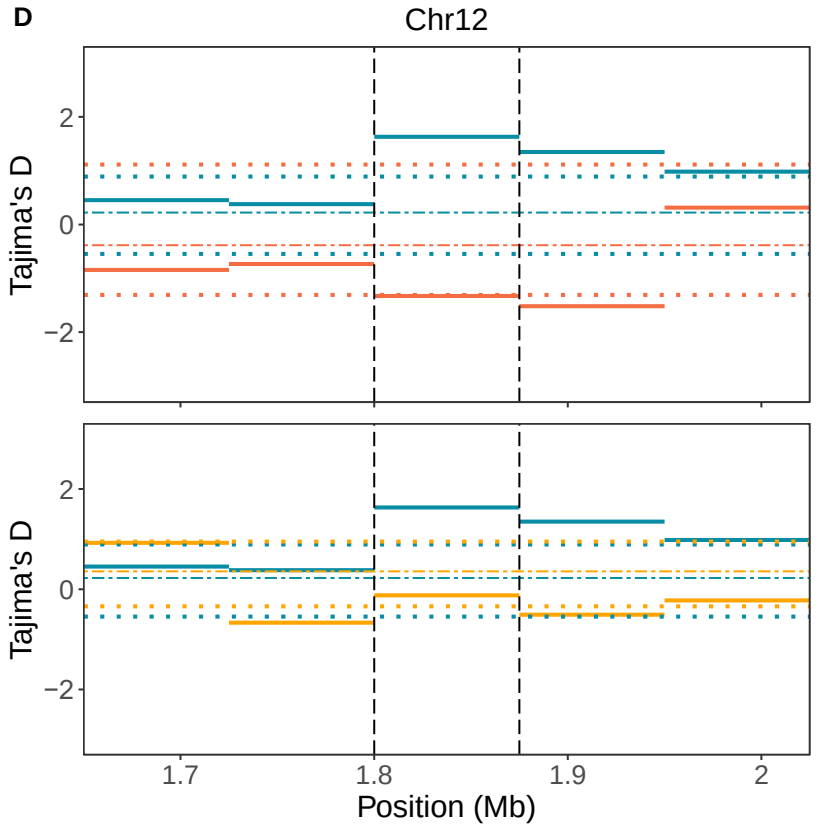
Table 3.6 Tajima's D in the overlapping outlier windows between GHP/GLP and GHP/TULP

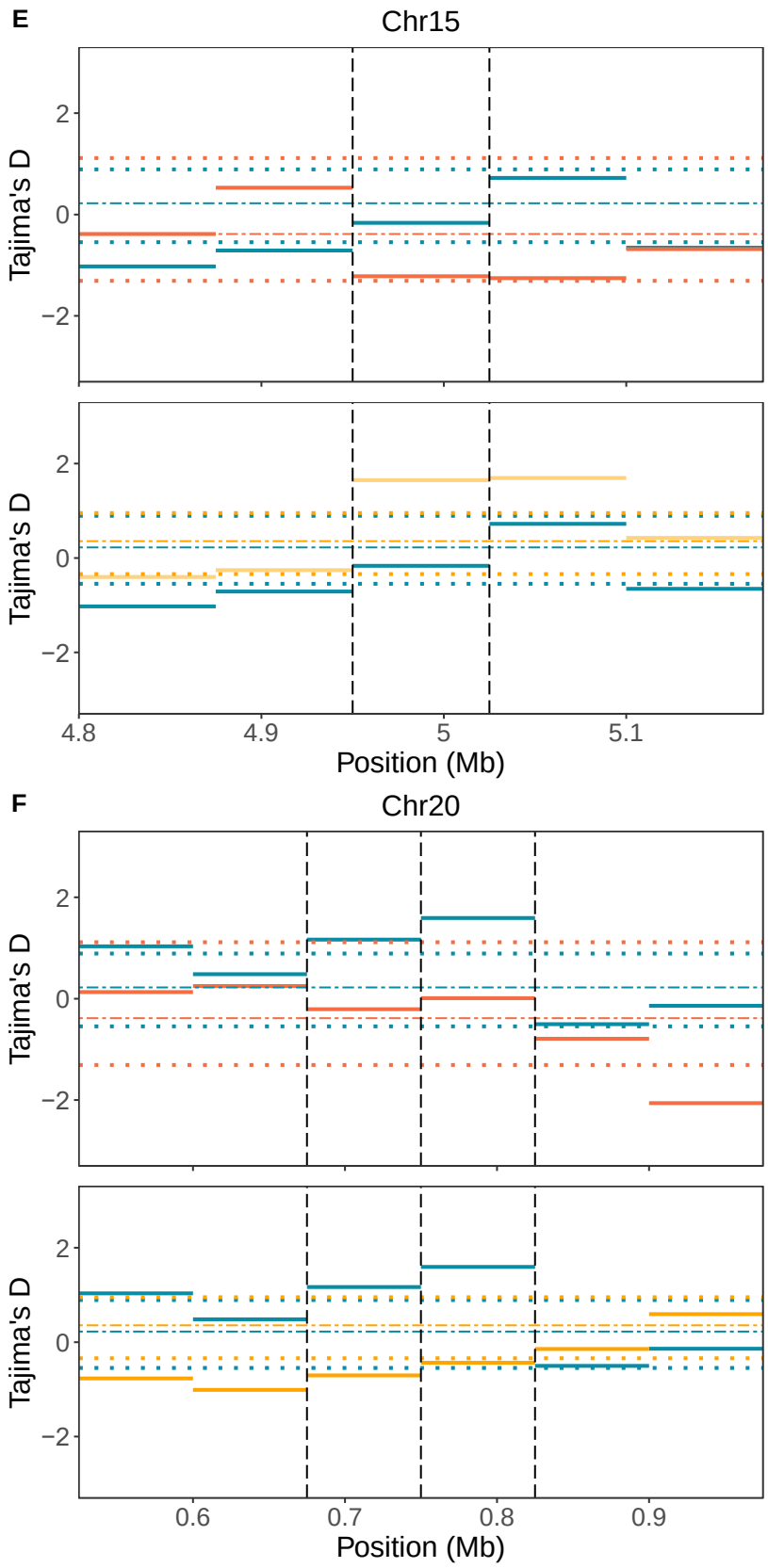
	Tajima's D			
Window	**GHP Median = 0.22**	**GLP Median = -0.38**	**TULP Median = 0.36**	**Selection in GHP, TULP or both?**
chr1:21450001-21525000	1.782	-1.964	-1.370	Both
chr6:18075001-18150000	1.450	2.802	1.027	GHP
chr10:4575001-4650000	1.916	-1.647	-0.487	GHP
chr12:1800001-1875000	1.631	-1.329	-0.121	GHP
chr15:4950001-5025000	-0.165	-1.217	1.647	TULP
chr20:675001-750000	1.166	-0.208	-0.705	Both
chr20:750001-825000	1.593	0.009	-0.437	GHP
000094F_0:750001-825000	1.204	-0.901	-0.532	GHP
000094F_0:975001-1050000	1.632	-0.254	-0.334	GHP
000094F_0:1275001-1350000	1.694	-0.308	-0.322	GHP
000094F_0:1575001-1650000	2.045	-0.189	-0.430	GHP
000113F_0:675001-750000	1.700	-0.507	-1.250	Both

When examining the intersect between GHP/GLP and ECHP/ECLP outlier lists, we found fewer convergence candidates; seven overlapping outlier windows across four chromosomes (table 3.5). One of the windows on chromosome 12, is the first of the four consecutive windows found in the within-river analysis for ECLP (figures 3.12 & 3.13), and when considering per SNP values of the measures we saw that outliers are evenly spread across the four genes in this window: laminin subunit gamma-3-like (LAMC3), allograft inflammatory factor 1 (AIF1L), 5'-AMP-activated protein kinase catalytic subunit alpha-1 (PRKAA1) and E3 ubiquitin-protein ligase TTC3 (TTC33) (figure 3.13).


A
Chr1
Tajima's D
2
0
−2
21.4
21.5
21.6
Position (Mb)
B
Chr6
18
18.1
18.2
18.3


C
Chr10
Tajima's D
2
0
−2
4.5
4.6
4.7
4.8
Position (Mb)


D
Chr12
Tajima's D
2
0
−2
1.7
1.8
1.9
2
Position (Mb)


E
Chr15
Tajima's D
2
0
−2
4.8
4.9
5
5.1
Position (Mb)
F
Chr20
Tajima's D
2
0
−2
0.6
0.7
0.8
0.9
Position (Mb)

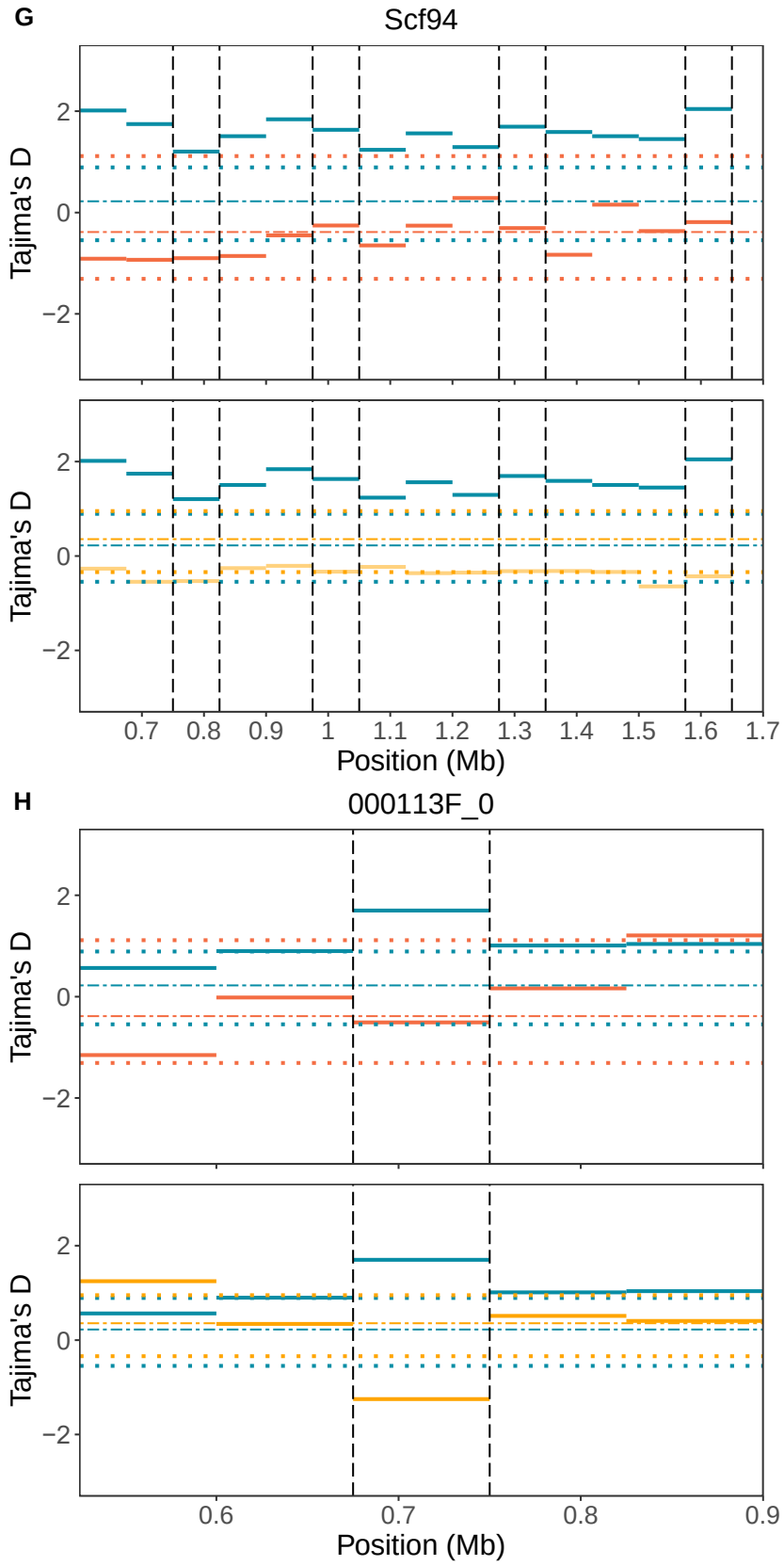


Figure 3.17 Tajima's D for GHP and the two LP populations in the identified outlier windows. Dark blue segments indicate values for GHP, orange segments indicate GLP, and yellow segments TULP. The outlier windows are located between the vertical dashed lines, and horizontal dashed lines the LP medians +/- 1 the median absolute deviation (MAD)(colours match populations).

3.5 Discussion

In Chapter 2 we described the levels of genetic variation in two experimental populations of guppies, their HP sources and a naturally colonised LP population. We found limited genetic variation in GLP, ECHP and ECLP, compared to high levels of variation in GHP and TULP, resulting from differences in their demographic histories. Here, we used whole genome sequencing data to investigate how these differences in the amounts of founding genetic variation affect selection within a river, and how they affect the likelihood of convergent molecular evolution among the population pairs. We found population specific signals of selection in all three pairs despite very low amounts of SGV in GLP and ECLP. Overlap among outlier windows using non-haplotype-based methods and the extended haplotype homozygosity methods in the introduced populations revealed that they experienced recent directional selective sweeps. On the other hand, there was little overlap between the non-haplotype-based methods and extended haplotype homozygosity method in GLP, indicating it is rare for variants in this population to have experienced recent selection that was strong and persistent enough to be driven to fixation (Coop *et al.*, 2009). Further, we show that both introduced populations have signatures of convergent evolution with the natural LP population, but found little evidence for convergence between the two experimental populations, indicating that each introduced population appears to have taken a different molecular approach to adapt to a low predation environment.

Differences in the demographic histories of the two experimental populations could lead to differences in selection, and could also affect our ability to detect selection using outlier approaches. When comparing the number of overlapping windows among measures within each river, we found that ECHP/ECLP

consistently had more overlapping windows than GHP/TULP (figures 3.2 & 3.3). This could be due to strong migration between ECHP and ECLP, and the lack of migration between GHP and TULP (see Chapter 2). Whiting & Fraser (2020) showed that in scenarios with migration, correlations between genome scan outliers were stronger than in scenarios without migration. In the presence of migration, levels of differentiation decrease in neutral loci, whilst population divergence around adaptive loci is retained, a pattern which increases the informativeness of measures with regards to their correlation with selection and in identifying outliers (Whiting and Fraser, 2020).

When we investigated within-river selection in ECHP/ECLP, we found over half of the overlapping windows were located on chromosomes 8, 10 and 12. Within the region on chromosome 12 we highlighted PTGES and PTGER4 as interesting genes. Both are prostaglandin E related genes. Prostaglandins are involved in the induction of ovulation and parturition in guppies (Venkatesh *et al.*, 1992). Female guppies in LP locations have their first brood at a later age than HP females, and they give birth less frequently (e.g. Reznick and Endler, 1982; Reznick, 1989), suggesting a role for prostaglandin and its related genes in the adaptation to LP environments. A third gene of interest in this region was USP20, which has been found to be involved in cranial skeleton development in zebrafish, *Danio rerio* (Tse, 2017): USP20 morphants had a shorter jaw structure and rounded Meckel's cartilage. Guppies from HP and LP environments have previously been shown to have different head morphologies, where LP guppies had shorter and rounder heads than HP guppies (Torres-Dowdall, Handelsman, Reznick, *et al.*, 2012).

Further investigations into patterns of heterozygosity in blocks of consecutive overlapping outliers on chromosome 8 and chromosome 12, revealed an extreme

loss of heterozygosity in these windows, especially in males. Chromosome 12 harbours the sex-determining locus (SDL) (Tripathi *et al.*, 2009), and therefore differences between the sexes could be expected on this chromosome, but why we also see these differences in autosomal regions is unclear. Differential natural selection between the two sexes can contribute to reduced observed heterozygosity in one sex, especially if it operates on traits that differ between the sexes, such as survival, growth or fertility (de Vries and Caswell, 2019). It is also possible that due to the high level of inbreeding in ECLP (see Chapter 2) only a few males contributed to the next generation, if females were choosing to mate with unrelated males in an attempt to avoid further inbreeding. Fitzpatrick and Evans (2014) provided evidence for postcopulatory inbreeding avoidance in captive-bred guppies, where inseminated sperm from unrelated males outcompeted inseminated sperm from a full sibling. There is mixed evidence for precopulatory inbreeding avoidance in guppies, however. Daniel and Rodd (2016) found that females guppies discriminate against kin or familiar males if they were previously mated, and Speechley *et al*. (2019) showed that female guppies increased their propensity to mate with multiple males in response to an increased risk of inbreeding, like following mating with a brother. Indeed, avoidance of inbreeding is proposed to be the primary mechanism driving female mate preference leading to negative frequency dependent selection (NFDS) in guppies (Hughes *et al.*, 2013). Considering we see a described pattern of heterozygosity loss in males in ECLP but not in TULP, an outbred population, this latter scenario of inbreeding avoidance seems a possible explanation. However, further investigations would be needed to elucidate the origin of sex-based differences.

In the natural population pair GHP/GLP, overlapping outliers among the non-haplotype-based methods did not have extended haplotype homozygosity, suggesting selective sweeps are more ancient. This is to be expected if selection occurs early on in the adaptation process and has since experienced a relatively stable environment, as the signal of a selective sweep will dissipate over time (Przeworski, 2002; Kim and Nielsen, 2004). Furthermore, any identified overlapping outliers within a single river need to be interpreted with caution, as they cannot be distinguished from the effects of drift and bottlenecks (Beatty, 1984). This is especially important in populations like GLP with a low number of founding individuals, small current effective population sizes and limited migration reaching the upper parts of the river, as these aspects make it more likely selection is overshadowed by the effects of random genetic drift (Fraser *et al.*, 2015).

To assess the occurrence of convergent molecular evolution among the three rivers, we compared overlapping outlier windows among the populations. The two introduced populations did not overlap with each other more than would be expected by chance in three of the four measures used. This lack of molecular convergence could have many reasons. For example, if selection occurred via soft sweeps; the genome scans employed here have reduced power to detect the modest changes in allele frequencies characteristic for soft sweeps (compared to the large changes observed in classic hard sweeps) (Hermisson and Pennings, 2005; Pritchard and Di Rienzo, 2010). Similarly, if adaptation has a polygenic basis it is expected that selection results in small allele frequency changes at numerous loci, making it difficult to detect them (Pritchard *et al.*, 2010). Evidence for convergent evolution of large effect loci have been found in this system before,

however, in both natural populations (Whiting *et al.*, 2020) and introduced populations (Fraser *et al*. 2015).

These differences in repeatability could stem from variation in shared SGV. If two populations share a substantial amount of SGV, molecular convergence is more likely to occur (Conte *et al.*, 2012), but if this shared SGV has been removed in one population, due to a bottleneck for example, molecular convergence becomes limited. Indeed, the differences in the evolutionary pasts of the two introduced populations in the current study might be the reason behind the lack of convergence between them. Some of the windows we identified as potential adaptation candidates in GHP/GLP and GHP/TULP have also been identified as regions of interest in other guppy populations in this system (Fraser *et al.*, 2015; Whiting *et al.*, 2020), but lack signatures of selection in ECHP/ECLP. For example, the region on chromosome 15 (4,950,000-5,025,000 bp), was highly heterozygous in GHP, whereas ECHP was fixed for this window (data not included), and consequently when introduced to ECLP had no variation in this region for selection to act on. Overall, we can conclude that convergent phenotypes found in the two experimental populations of guppies were not the result of selection on the same genomic regions. It is possible convergence occurs at other levels of biological organisation, for example different genes under selection could be part of the same genetic pathway. It would therefore be interesting to also include analyses investigating other levels of convergence in future examinations of these populations.

Although the two introduced populations showed little convergence, both were found to share outlier windows with the naturally colonised population, providing strong candidates for locally adapted genomic regions. We already described the outlier window on chromosome 15 as a strong candidate for convergent

adaptation between GHP/GLP and GHP/TULP. Within this region are genes belonging to the cadherin gene family, known to be involved in cell-cell adhesion and pigment cell migration (Fukuzawa & Obika, 1995; Nishimura *et al*., 1999), which could play a role in the colour differences between HP and LP males. Another strong candidate region for convergence between GHP/GLP and GHP/TULP is the six overlapping outliers on chromosome 20/scaffold 94. This region was also found as a region of interest for convergent evolution by Whiting *et al*. (2020). They describe two haplotypes stretching across the first 3Mb of the chromosome in the Caroni drainage, but rivers in the Oropuche drainage lack this haplotype. Our Tajima's D results suggest selection in this region largely occurred in GHP rather than TULP, with GHP having higher Tajima's D then the genome median. Investigating individual haplotypes in this region, we found that TULP (based in the Oropuche drainage) lacked the haplotype structure observed in the populations from the Caroni (GHP, GLP, ECHP and ECLP, data not included), consistent with the study by Whiting *et al*. (2020). This could be due to introgression between TUHP and TULP in this region, but specific analyses would be needed to investigate this. These two regions are promising subjects for local adaptation to HP/LP environments in the Trinidadian guppy system, and a next step in elucidating this process could be a functional analysis of the genes in these regions.

By comparing the process of adaptation in two experimentally introduced populations with contrasting evolutionary histories and different levels of neutral genetic diversity, we found signals of recent selection in both populations. The different levels of SGV in the two populations likely affected the incidence of molecular convergent evolution in this system. With a reduction of shared SGV due to historical bottlenecks in ECHP and ECLP, these population likely explored

alternative molecular pathways to achieve the same LP phenotype compared to other guppy populations in the Northern Range mountains of Trinidad. These results show the importance of incorporating demographic analyses in studies investigating the likelihood of molecular convergence.

Chapter 4 – INFERRING DEMOGRAPHIC HISTORIES OF RAPIDLY ADAPTING POPULATIONS

4.1 Abstract

Phenotypic evolution can occur more quickly than previously thought but understanding the genetic basis of rapid evolution is still in its infancy. Until recently it was difficult to detect such rapid adaptation at the genomic level. We use an *in situ* experiment in guppies to examine rapid evolution. Guppy populations that were transplanted from high-predation (HP) to low-predation (LP) environments have been shown to mimic naturally-colonised LP populations phenotypically in as few as 8 generations. In this chapter we analyse patterns whole genome variation in four populations recently introduced into LP sites along with the corresponding HP source population and a naturally evolved LP population. Four years after the introduction, we found only minor genome-wide genetic changes in all experimental sites when compared to the HP source. Based on runs of homozygosity, we found limited evidence for bottlenecks and inbreeding in 3 of the 4 introduced populations, whereas the fourth population exhibited patterns consistent with inbreeding and bottlenecks in the recent past. Furthermore, we also show that demographic inference using the coalescent was not possible for these recent populations, most likely because not enough time has passed for sufficient population divergence in allele frequency to occur between the HP source and the transplanted populations.

4.2 Introduction

Evolution was long thought to happen on a long timescale, one that could not be observed in real time (Gillespie, 1991). However, examples across different

taxa and traits (e.g. melanism in the peppered moth (van't Hof *et al.*, 2011), morphology changes in *Anolis* lizards (Losos, 2009) and life history of guppies (Endler, 1980)) have shown that evolution can be so fast it occurs on an ecological timescale. Rapid adaptation often happens in response to natural or human-mediated shifts in the environment (Reznick, Losos, *et al.*, 2019). In light of our fast-changing world, it is becoming increasingly important to understand if and how populations can cope with sudden changes in their environment, as it could help develop future conservation efforts tailored to a specific species or populations and help predict a population's response to future changes in their environment.

Sudden shifts in the environment, or colonisation of a new area, are often characterised by demographic events such as bottlenecks, migration, inbreeding and periods of exponential growth. These processes can leave traces in the genome that resemble selection, for example both selective sweeps and population bottlenecks can result in reduced genetic diversity (Tajima, 1989a), making it difficult to distinguish selection from demographic processes. When searching for candidates of selection, it is therefore important to incorporate demographic analyses to account for processes other than selection that may have shaped patterns of genetic diversity in a population.

In Chapter 2, I used two approaches for inferring demographic histories from genomic data: the first method is based on the site frequency spectrum (SFS), which could involve Bayesian computation (e.g. ABC, Beaumont, Zhang and Balding, 2002), or coalescent simulation (e.g. fastsimcoal2, Excoffier *et al.*, 2013). The second method is based on haplotype patterns (Beichman *et al.*, 2018), and analyses the abundance and length of homozygous tracts across the

genome that have arisen from limited effective population sizes (N_E) or through consanguineous mating (Ceballos *et al.*, 2019).

Methods based on coalescent modelling have mostly been used in well-established populations and it is unclear whether they can accurately infer very recent demographic histories (e.g. <20 generations). To investigate the initial process of rapid adaptation and the accuracy of current methods to infer the recent demographic history would require a population coming from a known source, and a known time of divergence, as this would allow to ground truth the models. Field experiments with wild populations, where the environment is intentionally manipulated to investigate the evolutionary response of a population, could therefore be used to test models of inference.

In Chapter 2, we successfully applied demographic inference methods to two long-term experimental populations of guppies in the Northern Range mountains of Trinidad (Haskins *et al.*, 1961; Reznick and Bryga, 1987), and found our demographic models for these populations supported previously known information about the timing of the introduction and the occurrence of bottlenecks in these populations. This shows these methods work for relatively young, but established populations (>60 generations), but it remains unclear whether these methods would also be capable of inferring more recent demographic histories of just a few generations (<20 generations).

A new guppy experiment

In 2008 and 2009, a new guppy introduction experiment was set up in the Northern Range mountains of Trinidad to investigate the early adaptation to LP environments. Guppies were transplanted from a single HP source in the Guanapo river (“GHP”) to four, previously guppy free, LP environments.

Additionally, in each introduction year, the canopy of one of the localities was thinned (Upper Lalaja and Taylor in for 2008 and 2009, respectively), which increased the primary productivity of these streams (Kohler *et al.*, 2012). The methods of sampling from the wild population and introducing fish to the new localities were the same for each population. In brief, a large sample of guppies were collected from GHP independently in each year and reared to maturity in single sex groups. Fish were subsequently mated in groups of five males and five females. Males and females from the same breeding group were released into different streams in the same year (i.e. females introduced into IT and IC came from the same breeding stock). Guppies possess receptacles that can store sperm (Lopez-Sepulcre *et al* 2013), meaning the introduced females carried sperm from the males of their breeding group when they were introduced with a new group of males in each respective river. This experimental approach was designed to increase the effective population size (N_E) of the populations at the time of introduction. In 2008, 38 fish of each sex were introduced to the Lower Lalaja ("ILL") and Upper Lalaja ("IUL") sites, each. In March 2009, 52 fish of each sex were introduced into the Taylor ("IT") site and 64 fish of each sex to the Caigual ("IC") site (figure 4.1, Arendt *et al*., 2014).

Phenotypic investigations have shown these populations have experienced selection since the introduction. Kemp *et al*. (2018) found that all populations, except for Caigual, diverged from the GHP source phenotype in coverage of blue/green, and in the case of IUL and IT, a reduction in black. Reznick *et al*. (2019) showed introduced male guppies evolved to mature at a later age and larger size in all four populations compared to GHP, but the onset of this evolution happened only after populations reached their peak densities, which differed between the introduction years. ILL and IUL were introduced with much lower

initial densities (due to less introduced fish and larger stream size than IC and IT) and it took three years for them to attain peak density (Reznick, Bassar, *et al.*, 2019). IC and IT, on the other hand, had high initial densities and reached peak densities after just two years. In addition to overall population growth, all four populations experience seasonal fluctuations, with populations decreasing during the rainy seasons. Populations with thinned canopies (IUL and IT) experienced more severe fluctuations in population densities compared to ILL and IC. These fluctuations were particularly clear in IT, where Reznick *et al*. (2019) observed a crash of the population density in the first year that nearly caused the extinction of the population.

4.3 Methods

Sampling and data generation

Fish from GHP, GLP, and the four introduction experiments (ILL, IUL, IC, and IT) were sampled in the spring of 2013. From each sample site, approximately 20 individuals were collected (N= 94, table 4.1). Fish were stored in 95% ethanol and stored at -20°C. DNA was extracted from caudal peduncle tissue using the Qiagen DNeasy Blood and tissue kit. Individual samples were prepared for whole

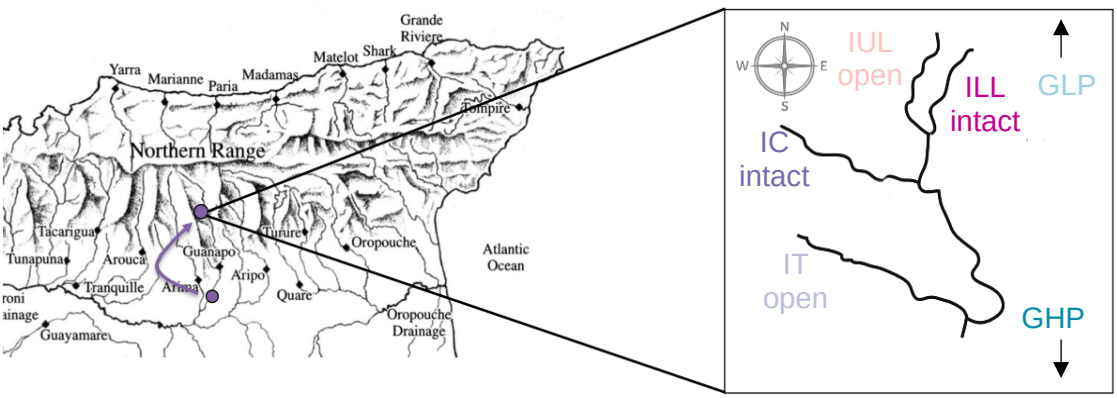


Figure 4.1 Map showing the locations of the introduction experiments. GHP is located further downstream and GLP further upstream. Open/intact indicates the treatment of the canopy.

genome sequencing following the Illumina TruSeq DNA sample preparation guide with approximately 250bp insert size. Eighty-six samples were sequenced by multiplexing 6-10 individuals per lane on an Illumina HiSeq 2000 and 3000. The remaining eight samples were sequenced using the Illumina HiSeq 4000 with a 150bp paired end read metric.

Read mapping

We followed the same protocol as described in Chapter 2 to process raw sequence reads. Briefly, we assessed the quality of the reads with FastQC (Andrews, 2010; Puritz *et al.*, 2014) and adapters and low-quality bases were removed with TrimGalore! (Krueger, 2012). Trimmed sequence reads were processed following GATK v4.0 Best Practices protocol. A more detailed description of this protocol can be found in Chapter 2. Resulting variants were then further filtered in VCFTOOLS (V0.1.12b) (Danecek *et al.*, 2011): only loci with a minimum depth of 5x and a maximum depth of 200x were kept. The VCF was split by population and the population files were filtered for sites where at least 50% of the individuals were genotyped. Finally, the population files were merged again and filtered for a minor allele frequency of >0.01. The final VCF file contained 5,700,890 SNPs.

Table 4.1 Summary of population information.

Population	ID	Year	Canopy treatment	N	Males	Females	Mean per individual raw reads	Mean per individual clean reads	Mean depth
Lower Lalaja	ILL	2008	Intact	13	3	10	92,006,734	87,470,908	11.6 x
Upper Lalaja	IUL	2008	Open	15	11	4	83,882,866	69,732,343	7.6 x
Caigual	IC	2009	Intact	15	11	4	62,413,477	57,276,037	7.2 x
Taylor	IT	2009	Open	14	8	6	63,813,611	60,794,284	8.1 x
Guanapo high predation	GHP	NA	NA	19	9	10	84,794,541	80,947,279	10.4 x
Guanapo low predation	GLP	NA	NA	18	10	8	99,093,373	94,135,861	12.1 x

For the haplotype analyses, we followed the double-phasing protocol as described in Malinksy *et al*. (2018): population VCF files were first phased per chromosome using BEAGLE (Browning and Browning, 2007), followed by a second round of phasing with SHAPEIT2 (Delaneau *et al*., 2012). SHAPEIT2 has an increased accuracy compared to BEAGLE (Delaneau *et al.*, 2012), but does not accept missing data. Therefore we use BEAGLE, which does accept missing data but has a high switch rate error (Delaneau *et al*., 2012), to create pre-phased VCF files that can be used as input for SHAPEIT2.

Neutral diversity and population structure

Population specific summary statistics (nucleotide diversity (π),Tajima's D and global F_{ST}) were calculated with PopGenome (Pfeifer *et al.*, 2014) and VCFTOOLS (expected heterozygosity, H_e).

A principal component analysis (PCA) was performed to assess population structure. To ensure the computed PC's reflect genome wide structure and not local linkage disequilibrium (LD) patterns, the VCF was pruned by running –indep-pairwise in PLINK (v1.90b6.7, Chang *et al*., 2015) with window size set to 50 bp, step size to 5bp, and the pairwise r^2 threshold to 0.2, after which eigenvectors and eigenvalues were calculated in PLINK using --pca.

Runs of homozygosity

The distribution and frequency of runs of homozygosity can be used for insight into the evolutionary history of individuals and populations. The presence of a few long ROH in a population indicates it recently experienced inbreeding or a bottleneck, whereas the presence of many, but very short ROH are indicative of a more ancient event, as recombination will have broken up ROH from more

distant ancestors over time (Kardos *et al.*, 2017). ROH in an individual are the result of family relatedness between its parents, and since some individuals will have consanguineous parents in highly inbred populations, this will result in greater variation in the frequency and total length of ROH exist than fully outbred populations (Kirin *et al.*, 2010). ROH can also provide an insight in to relative population sizes, with small populations containing more ROH compared to larger populations as a result of increased inbreeding (Ceballos *et al.*, 2018).

To assess levels of homozygosity and inbreeding, runs of homozygosity (ROH) were calculated for each individual following the same protocol as described for the established introduction populations in Chapter 2. Briefly, to minimize the number of ROH that occur by chance, the minimum number of SNPs a ROH must contain was calculated for each population following Purfield *et al*. (Purfield *et al*., 2017):

$$l = \frac{\log_e \frac{\alpha}{n_s * n_i}}{\log_e\left(1 - \overline{het}\right)}$$

Where n_s is the number of SNPs per individual (5,433,691 without the sex chromosome), n_i is the number of individuals, α is the percentage of false positive ROH (set to 0.05), and $\overline{\text{het}}$ is the mean heterozygosity across all SNPs. Each run had to be at least 500 kilobases long to exclude short, common ROH present in all individuals and all populations. Finally, at most 1 heterozygous site per window was allowed. Runs of homozygosity were estimated for each individual separately, and resulting ROH were binned into 0.5-0.75 Mb, 0.75-1.0 Mb, 1.0-1.5 Mb, and >1.5 Mb. To calculate the genomic inbreeding coefficient F_{ROH}, we used:

$$F_{ROH} = \frac{\Sigma L_{ROH}}{L_{auto}}$$

Where ΣL_{ROH} is the total length of an individual's ROH above a specified length threshold and L_{auto} is the length of the autosomal genome.

Demographic history analysis

We used the same method described in Chapter 2 to obtain a folded 2D SFS that also contains sites that are monomorphic across the populations. For a detailed description of this method, see Chapter 2. Briefly, we created a VCF file with only monomorphic sites using GATK v3.8 (van der Auwera *et al.*, 2014), which was then split by population and filtered for sites where at least 80% of the individuals were genotyped to minimise missing data. Pairwise HP/LP VCF files were created for GHP and each of the LP populations by merging the GHP VCF file and the LP VCF files with CombineVariants (GATK, v3.8). Next, we created population VCF files containing SNPs from the VCF file used in the other analyses by splitting the original VCF by population and applying the same missing data filter of 80% to each file. Pairwise HP/LP SNP VCF files were created the same way we created the monomorphic pairwise VCF files. Finally, the SNP VCFs and monomorphic VCFs were pruned and subsequently merged using GATK's (v4.0.4.0) MergeVcfs, and SFS were generated with EASYSFS (Overcast, 2017).

Model choice

Four different demographic models were fitted to the observed 2D SFS of each population pair using FASTSIMCOAL2 (Excoffier *et al.*, 2013). We ran each model with and without migration, resulting in a total of eight models per populations (figure 4.2). The different models estimate effective population sizes (N_E) and whether or not a bottleneck occurred in the introduced populations. Models 1 and

3 allow for population growth or contraction in the introduced population, while GHP remains stable (model 1) or also experiences growth or a contraction (model 3). Models 2 and 4 include a bottleneck in the introduced population that can occur at any time after the introduction event, while GHP remains stable (model 2) or experiences growth or a contraction (model 4). As for the models described in Chapter 2, the model was unrestricted and free to determine whether a population experienced growth or a contraction. The initial parameter values of each run were drawn from set ranges (supplementary table S4.1). The simulations were run with the same initial parameter ranges (supplementary table S4.1). Only the initial values were drawn from these ranges, after that the upper boundary does not imply a hard limit for the parameter range and the model can surpass this number by 30% in each cycle (fastsimcoal manual).

Each model was tested by running 100 independent iterations of FASTSIMCOAL2 using the options: -n 200000 (number of simulations), -m (to estimate the SFS for the minor allele), -M (parameter estimation by maximum composite likelihood), and -L 100 (number of ECM cycles). We used a mutation rate of 4.8×10^{-8} (Künstner *et al.*, 2016) and a recombination rate of 4×10^{-5}. To determine which model best fits the demographic history of each population pair, we selected the run with the highest likelihood (the smallest difference between the observed and expected likelihoods) and used the parameters of this run to simulate 100 SFS to obtain a likelihood distribution for each model. These distributions were then compared to select the best model. Parameters from the best run were then used as initial values to simulate 100 independent SFS per population. By using the parameters from the best run, we can eliminate variance in the optimisation process and instead obtain evolutionary variance. Each simulated SFS was used in two independent runs of FASTSIMCOAL2, using the

same settings as above, resulting in 200 sets of parameter estimates which were then used to calculate medians and the 95% confidence intervals for each parameter.

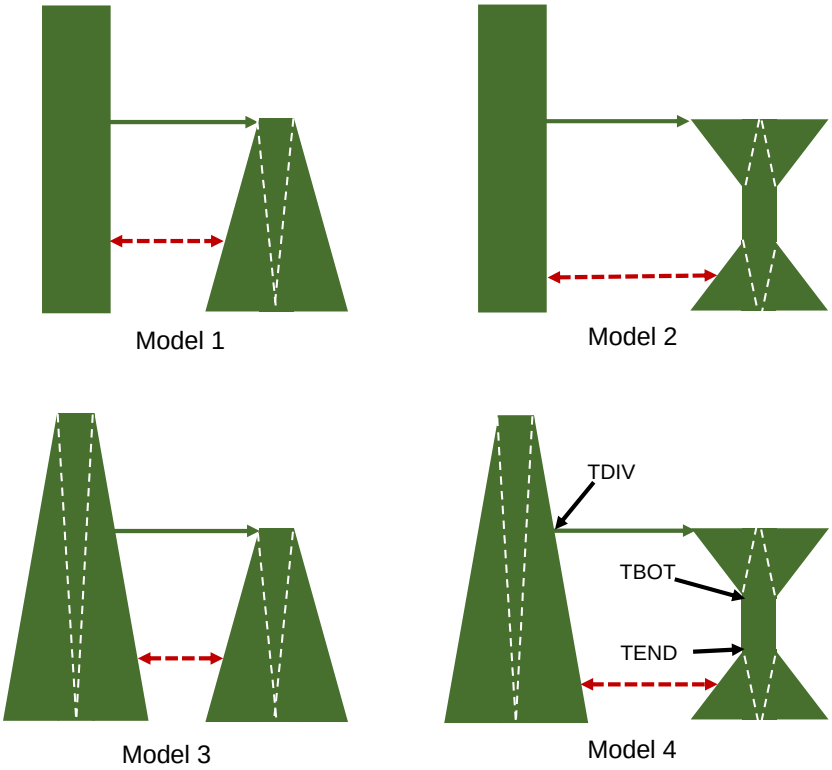


Figure 4.2 Schematics of the demographic models tested with FASTSIMCOAL2. Red lines indicate migration, white dashed lines indicate population growth or contraction.

4.4 Results

Whole genome sequencing data was generated for six populations of guppies from the Northern Range Mountains of Trinidad for a total of 94 individuals, ranging from 13 to 19 individuals per population (table 4.1). A total of 8.01 billion raw reads were generated, after filtering 7.4 billion clean reads remained (92%), with an average of 75.56 million reads per individual. Average coverage per individual across populations was 9.5x (min = 7.2x in IC, max = 12.1x in GLP, table 4.1). The final filtered dataset is a Variant Call Format (VCF) file with 5,700,890 SNPs.

Population structure

A PCA revealed little population structure between the introduced populations and their source on PC1, the introduction populations still largely cluster with GHP. The first axis (PC1) captured 21.15% of variation between the populations (figure 4.3a), separating GLP from GHP and the introduction populations. Along PC2 (6.46% of variation), GHP clusters more closely with IC and IT (the populations introduced in 2009 with high initial densities), whereas ILL and IUL (introduced in 2008 with low initial densities) are separated from this cluster. PC3 (5.56% of variation) again clusters IC and IT with GHP, but it separates ILL and IUL in opposite directions (figure 4.3b). Finally, we observe differences in the amount of variation within population. ILL and IUL show more variation than IC and IT. GLP had almost no variation at all, in line with results from Chapter 2, where we showed GLP had very low values of nucleotide diversity and expected heterozygosity.

Overall, pairwise F_{ST} was low between introduced populations and their source (median F_{ST} = 0.013 – 0.031) and was always high when paired with GLP (table 4.2). Of the introduced populations, fish from IT were more differentiated from GHP than fish from the other introduced populations were to GHP (table 4.2, $p<0.0001$ for IT/GHP vs ILL/GHP, IUL/GHP or IC/GHP, respectively, Mann-Whitney U test). Among experimental populations, comparisons with IT had in general higher values of F_{ST} than comparisons with ILL, IUL and IC had amongst one another, although all values were significantly different from each other (table 4.2, $p<0.0001$ for all comparisons, Mann-Whitney U test). IC was least differentiated from IT (median F_{ST}=0.027), and of the remaining pairwise comparisons between introduced populations, ILL and IUL had the lowest values

of F_{ST} (F_{ST}=0.017), indicating that guppies introduced in the same year are more closely related to one each other than guppies from other pairwise comparisons. In addition to the whole genome set, we investigated global F_{ST} values of the data set with outlier windows removed (supplementary table S4.2) and with the sex chromosome (chromosome 12) removed (supplementary table S4.3). Removing outlier windows resulted in a very minor average reduction of F_{ST} values of 0.005 for the median and 0.01 for the mean, and made no difference for the observed patterns among population pairs described above. Removing the sex chromosome made no difference for most values, and no significant difference for those that did change slightly: GLP/ILL (p=0.470, Mann-Whitney U test), IUL/IT (p=0.578), GHP/IC (p=0.848) and GLP/IT (p=0.933).

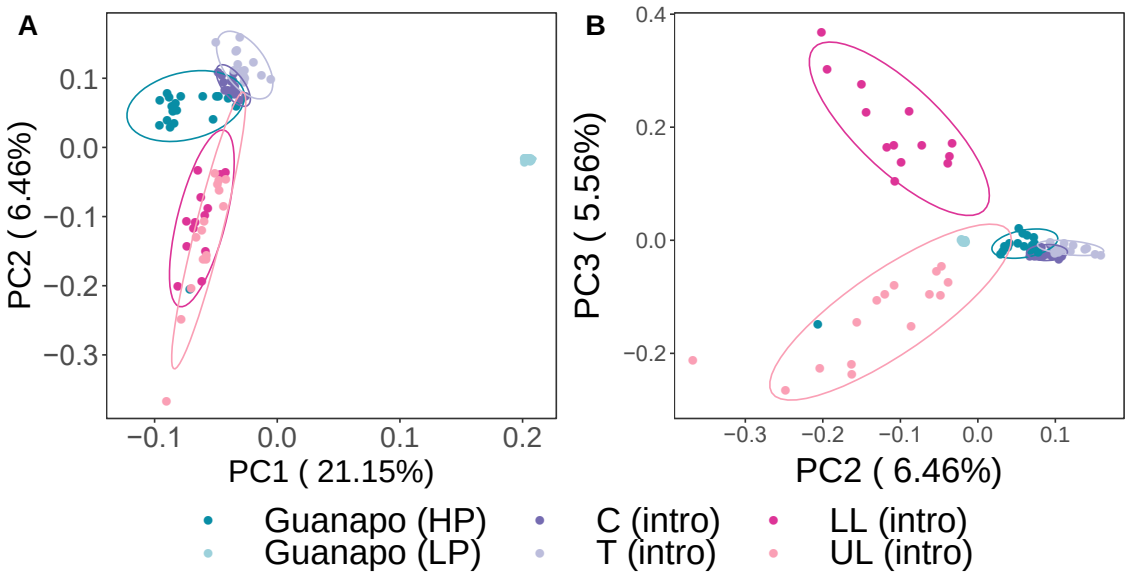


Figure 4.3 Principal component analysis of all six populations: A) Plotting PC1 (21.15% of variance explained) against PC2 (6.46% of variance explained). B) Plotting PC2 against PC3 (5.56% of variance explained). Ellipses represent the 95% confidence interval.

Table 4.2 Pairwise global F_{ST} values among all six populations. Mean F_{ST} below the diagonal, median F_{ST} above the diagonal. Population names can be found in table 1.

	ILL	IUL	IC	IT	GHP	GLP
ILL		0.017	0.024	0.031	0.013	0.267
IUL	0.028		0.022	0.031	0.013	0.271
IC	0.036	0.032		0.027	0.015	0.290
IT	0.044	0.042	0.040		0.022	0.281
GHP	0.025	0.022	0.024	0.035		0.309
GLP	0.282	0.286	0.305	0.297	0.317	

Neutral diversity

The introduction of guppies from GHP to ILL, IUL, IC and IT led to very minor changes in genetic diversity in the introduction populations (figure 4.4, table 4.3), whereas the natural LP population (GLP) experienced much larger changes. IC and IT experienced a significant decrease of H_e, with a 15.2% decrease in IT, and a 14.4% decrease in IC ($p<0.0001$ for all three, Mann-Whitney U test), whereas ILL and IUL experienced very slight increases of H_e in comparison to the HP source: 0.8% in IUL ($p = 0.0235$ Mann-Whitney U test) and 4.9% in ILL ($p < 0.0001$). Similarly, median nucleotide diversity (π) decreased by 15% in IT and 13% in IC, and was slightly, but significantly, increased in ILL and IUL (8% and 7% respectively, $p<0.0001$ Mann-Whitney U test). Tajima's D was shifted to a more positive value in IC and IT compared to GHP (figure 4.4, table 4.3, $p<0.0001$), suggesting a loss of rare alleles compared to the HP source, which is indicative of a sudden population contraction. In ILL and IUL Tajima's D has decreased slightly compared to GHP, but stayed positive (figure 4.4, table 4.3, $p<0.0001$), implying an increase of rare alleles which could be related to a recent selective sweep or a population expansion after a recent bottleneck. In GLP, Tajima's D was negative, the result of an excess of rare alleles resulting from a population expansion after a bottleneck, or a recent selective sweep.

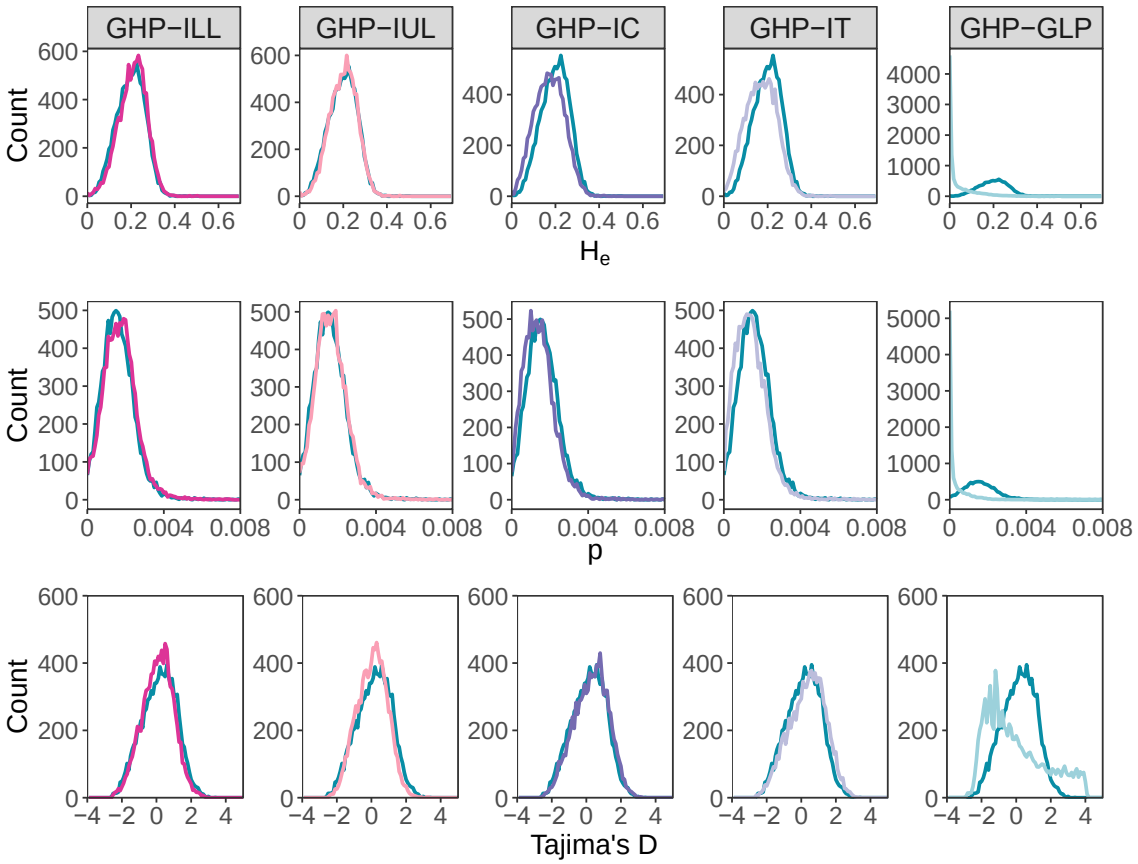


Figure 4.4 Frequency distributions of summary statistics, expected heterozygosity (H_e, top panel), nucleotide diversity (π, middle panel) and Tajima's D (bottom panel). ILL and IUL had a slightly increased H_e and π, and a reduced Tajima's D. IC and IT had reduced H_e and π, and an increased Tajima's D. Note the different y-axes for GLP in H_e and π.

Table 4.3 Summary of population genetic statistics. Mean and median values of expected heterozygosity (H_e), nucleotide diversity (π) and Tajima's D (TD)).

Population	Mean H_e	Median H_e	95% quantile H_e	Mean π	Median π	Min-max π	Mean TD	Median TD	min-max TD
ILL	0.207	0.212	0.076 - 0.319	0.00172	0.00168	0 - 0.0100	0.079	0.121	-2.531 - 3.023
IUL	0.200	0.203	0.077 - 0.313	0.00166	0.00166	0 - 0.0089	0.043	0.085	-2.594 - 3.340
IC	0.172	0.173	0.046 - 0.301	0.00143	0.00135	0 - 0.0079	0.274	0.340	-2.624 - 3.167
IT	0.170	0.171	0.039 - 0.302	0.00141	0.00132	0 - 0.0081	0.346	0.427	-2.629 - 3.358
GHP	0.198	0.202	0.072 - 0.314	0.00163	0.00156	0 - 0.0097	0.158	0.212	-2.612 - 3.376
GLP	0.035	0.006	0 - 0.202	0.00027	0.00004	0 - 0.0043	-0.023	-0.440	-2.865 - 4.097

Runs of homozygosity

We found ROH of > 0.5 Mb in 93 of the 94 individuals, with one individual in GHP lacking ROH fitting requirements we set to identify ROH. Results for GLP are presented in Chapter 2 and will not be further discussed here. Of the four introduced populations, IT had the most ROH (N=394, table 4.4), followed by IUL (N=163), ILL and IC had a similar amount of ROH (N=142 and N=149, respectively), and GHP had the lowest number of ROH (N=83). We also observed variation in the maximum length of ROH among populations. IT was the only population with ROH > 2Mb (excluding GLP, figure 4.5, table 4.4), suggesting it experienced recent inbreeding. ILL, IUL and IC also have more ROH and more of their genome covered in ROH when compared to GHP source, suggesting there may have been a limited bottleneck effect since the introduction, although not as severe as IT (figure 4.5 and table 4.4). The extremely low number of ROH, and the lack of ROH > 1.5Mb in GHP is indicative of a large and stable population, with potential admixture from other populations.

Table 4.4 The average length and number of ROH per individual in each population.

Population	Total # ROH	Mean Mb in ROH	Min-max length (Mb)	F_{ROH} (min-max)
ILL	142	7.4	0.50-1.97	0.010 (0.004-0.027)
IUL	163	7.5	0.50-1.65	0.010 (0.002-0.021)
IC	149	6.8	0.50-1.87	0.009 (0.003-0.037)
IT	394	21.4	0.50-2.77	0.029 (0.011-0.077)
GHP	83	2.95	0.50-1.21	0.004 (0.001-0.014)
GLP	4083	170.9	0.50-2.35	0.231 (0.097-0.322)

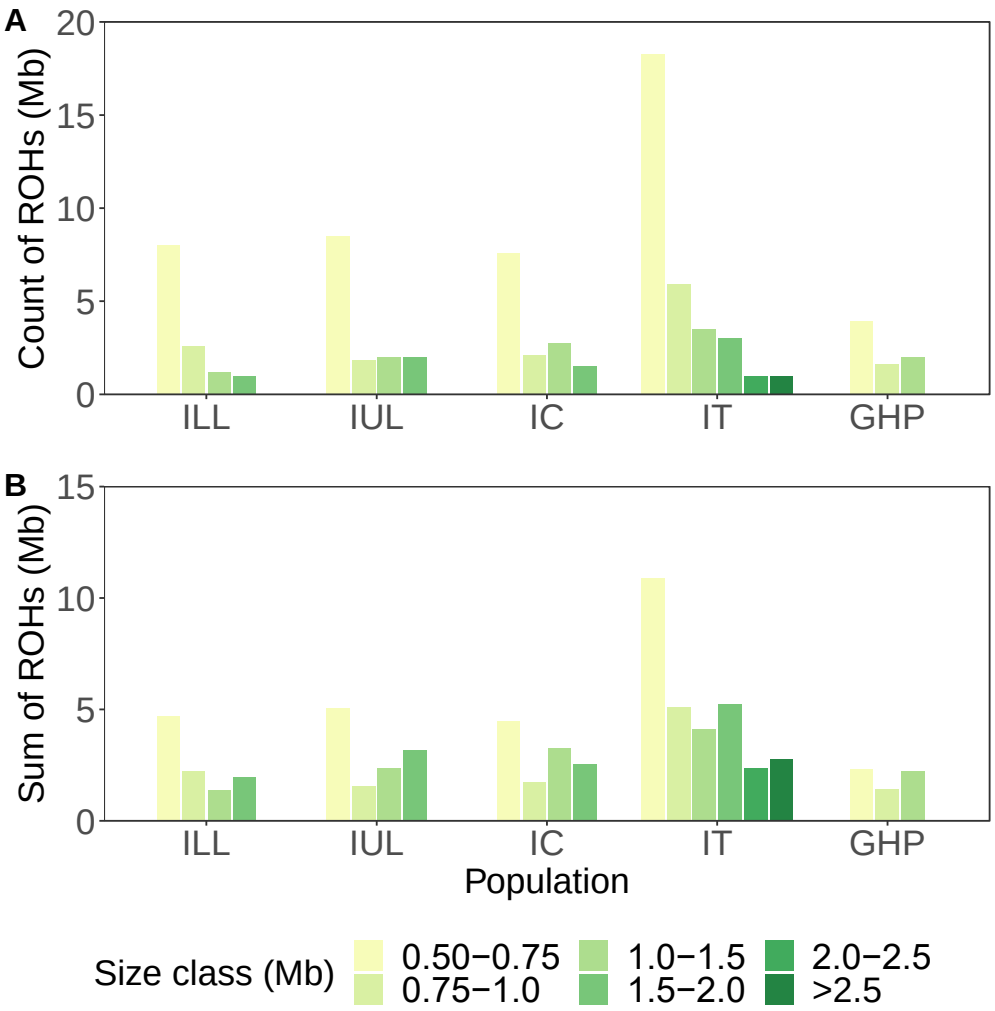


Figure 4.5 Plots showing ROH per size category. A) Mean number of ROH's per individual for each size category. B) Mean sum of ROH's per individual for each size category.

Inbreeding coefficients (F_{ROH}) revealed similar patterns between the introduced populations (figure 4.6). F_{ROH} in IT was significantly higher compared to the other populations, ranging from 0.011 to 0.077, with a median of 0.026 (table 4.4, $p<0.0001$, Mann-Whitney U test). This is further evidence IT experienced a period of reduced population size that resulted in increased inbreeding. Values of F_{ROH} for ILL, IUL and IC were not significantly different from each other (ILL-IC: $p=0.185$, ILL-IUL: $p=0.856$ and IUL-IC: $p=0.305$, Mann-Whitney U test). ILL, IUL and IC, however, had higher F_{ROH} values compared to GHP (GHP-ILL & GHP-IUL: $p<0.0001$, GHP-IC: $p=0.001$, Mann-Whitney U test), supporting the idea the introduction may have left a weak bottleneck signature in these populations.

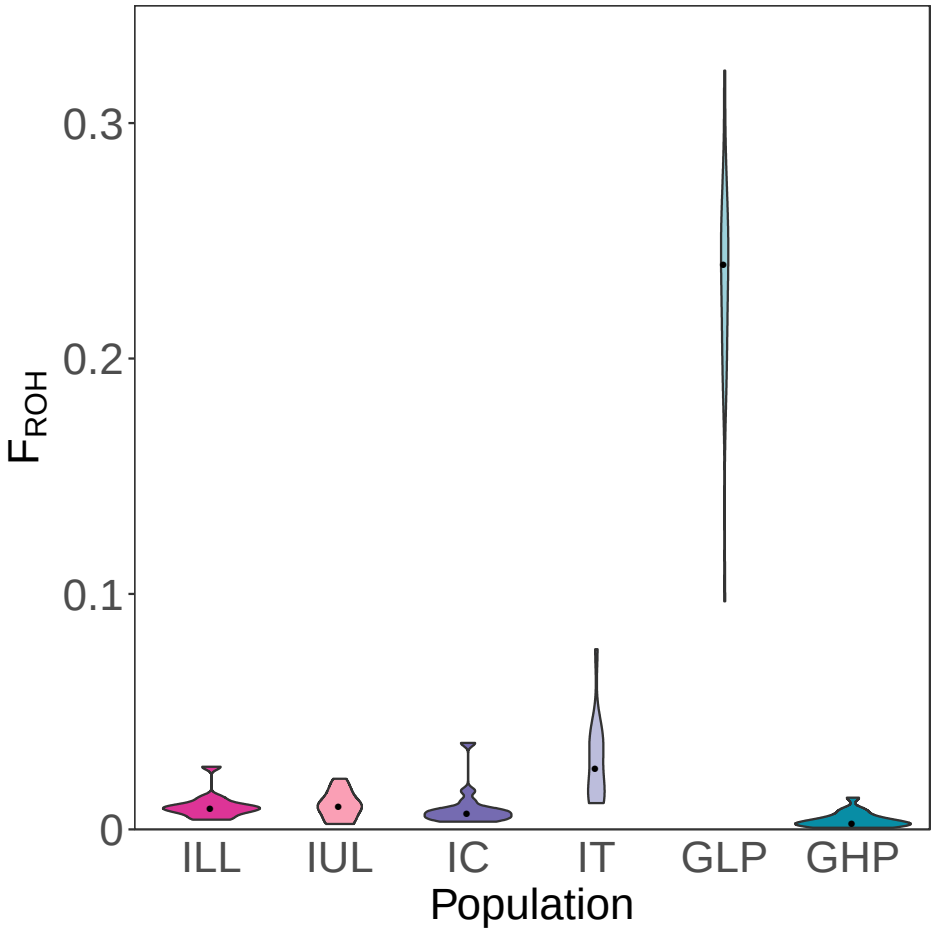


Figure 4.6 F_{ROH} per individual per population. F_{ROH} is the proportion of the autosomes covered in ROH.

Demographic history analysis

To investigate specific demographic parameters of each population's evolutionary history we used coalescent modelling to estimate effective population sizes, timing of any bottlenecks and migration rates. After running four different models with and without migration (a total of eight models per population), we could not identify a single model that had the highest likelihood for IUL and IC based on the likelihood distribution (figure 4.7b, c). For IT, model 4M has a slightly higher likelihood than model 2M (table 4.5), however there is no clear distinction between the two models. Finally, ILL was the only population with a clear most likely model (model 2M, figure 4.7a), but this model had large variability in the likelihood distribution, suggesting it might not accurately reflect the populations' history.

Indeed, when we ran the parameter simulations for this model for ILL, we found the point estimates from the best run did not always fall within the 95% confidence interval (supplementary table S4.4a), suggesting this model does not describe the demographic history of ILL and GHP well. Despite not being able to identify a best model for the remaining populations, we ran simulations for the model with the highest likelihood distribution (model 4M in all three populations, table 4.5). However, we found that the parameter estimates were not consistent with what is known about the populations, providing further evidence that the models do not accurately reflect the demographic histories of these populations (supplementary table S4.4. For example, estimated divergence time for IUL, IC and IT was overestimated (supplementary table S4.4, estimated time of split was 27, 71, and 86 generations respectively, actual time of divergence was 10 generations for IUL, and 8 generations for both IC and IT, assuming a generation time of 2 generations per year (Endler, 1980)). Whereas time of divergence was slightly

under estimated for ILL (estimated 7 generations versus 10 generations since the actual time of split). Furthermore, estimates of GHP effective population size, which should be similar among the pairwise models, varied extensively, ranging from 545 individuals in the IUL model to 11,608 in the IT model (supplementary table S4.4).

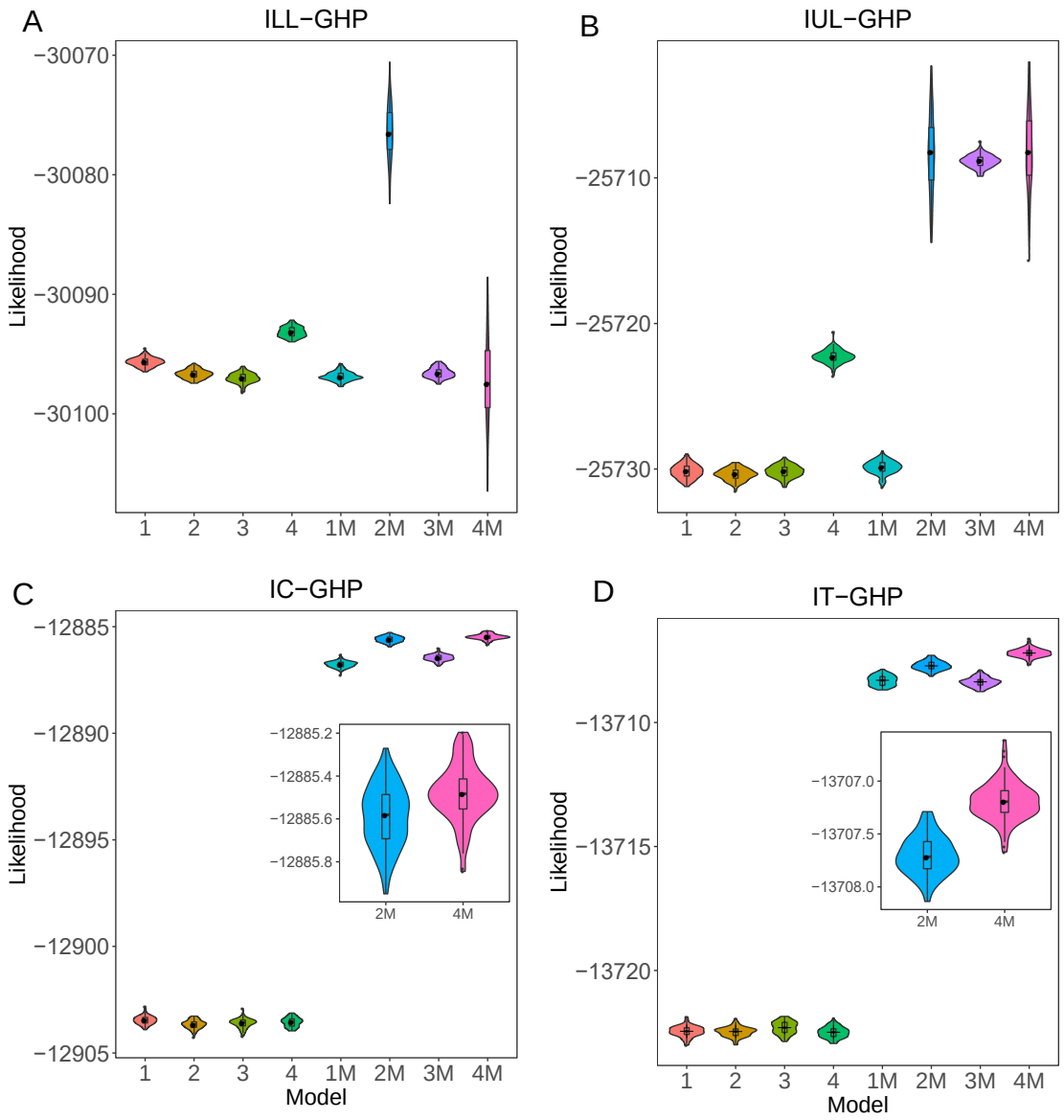


Figure 4.7 Likelihood distributions of each model for each HP-LP pair, obtained from the best initial runs in FASTSIMCOAL2. A) ILL/GHP models, B) IUL/GHP, C) IC/GHP models, and D) IT/GHP. Inserts in C and D picture the best two models in close up.

Table 4.5 Δ-likelihoods for the best runs of each model in FASTSIMCOAL2. These runs were then used to simulate a likelihood distribution of 100 SFS. The best models are printed in bold.

	No migration				Migration			
Population	**model 1**	**model 2**	**model 3**	**model 4**	**model 1**	**model 2**	**model 3**	**model 4**
ILL/GHP	-65.21	-66.38	-66.55	-63.06	-66.63	**-41.12**	-65.82	-56.45
IUL/GHP	-97.90	-98.09	-98.10	-90.53	-98.02	-69.17	-77.40	**-67.15**
IC/GHP	-26.92	-27.06	-26.89	-27.12	-10.37	-9.15	-9.90	**-9.09**
IT/GHP	-30.07	-30.34	-30.07	-30.02	-16.25	-15.85	-16.50	**-15.30**

4.5 Discussion

In this chapter, we set out to investigate patterns of neutral genomic variation and demographic histories in four replicate experimental populations of guppies. After four years since the introduction (8-10 generations ago), we observed minor genome-wide genetic changes in all experimental sites when compared to the GHP source. Based on the abundance and length of runs of homozygosity, we found evidence for bottlenecks and inbreeding was minimal in 3 of the 4 populations (ILL, IUL and IC), whereas one population (IT) exhibited patterns consistent with inbreeding and bottlenecks in the recent past. Finally, we found that demographic inference based on the SFS was not possible for these populations, most likely because not enough time has passed for sufficient population divergence in allele frequency to occur between the HP source and the transplanted populations.

Our results revealed evidence of a genetic bottleneck in IT, consistent with the population size crash observed by Reznick *et al*. (2019). We found that IT had the lowest levels of expected heterozygosity and nucleotide diversity, and the greatest increase in Tajima's D, all indicators of a recent bottleneck (Smith and Haigh, 1974; Tajima, 1989a). Furthermore, IT was the only population with ROH > 2Mb, and its F_{ROH} was significantly higher than those observed in the other three introduced populations. Compared to values of F_{ROH} in wild populations from other species it is considered inbred. For example, Hooper *et al*. (2020) consider populations of killer whales with $F_{ROH} < 0.015$ to be inbred. Because demographic inference of the introduced populations was not successful, we cannot identify the specifics of this bottleneck (e.g. timing, duration and bottleneck size). Reznick *et al*. (2019) however, observed a severe crash in the population density of IT during the first year, and two other studies (Torres-Dowdall,

Handelsman, Ruell, *et al.*, 2012; Fitzpatrick *et al.*, 2014) found IT had a higher mortality rate than IC in the first year, which they attributed to stronger influence of seasonal changes and indicators of disease in IT. Together, these results suggest that the bottleneck we detected probably occurred in the first year after the introduction, and is not the result of a founding bottleneck.

The remaining three populations mostly showed similar signals of constant population sizes. Even though they had slightly increased numbers and length of ROH compared to GHP, ILL, IUL and IC had fewer and shorter ROH compared to IT, consistent with a lack of recent inbreeding or bottlenecks. Indeed, Reznick *et al*. (2019) found that these three populations experienced continuous growth during the first three years of the experiment. Even during population declines in the rainy season, the population densities do not drop to the levels observed at the start of the experiment, whereas in IT during the first year the population density drops to the levels observed at the start of the experiment.

Reznick *et al*. (2019) showed that the onset of selection for age and size at maturity differed between the different initial density treatments. For all populations, selection started after they reached peak densities, however populations with low initial densities (ILL and IUL) took three years to attain this, whereas the high initial densities (IC and IT) started to evolve after just two years. We find evidence of these differences in population growth and contraction at the genomic scale. The low density populations, ILL and IUL, saw a slight increase in expected heterozygosity and nucleotide diversity and a reduced Tajima's D, which is indicative of a population that experienced rapid population expansion (Zenger *et al.*, 2003; Ortego *et al.*, 2007). In the high-density populations on the other hand, IC and IT, we found reduced expected heterozygosity and nucleotide diversity, and increased Tajima's D, a signal of less population growth.

To investigate the effect of increased primary productivity on adaptation in guppies one of the sites had their canopy thinned in each year (IUL and IT were thinned, while canopies in ILL and IC remained intact). Kemp *et al*. (2018) reported that fish from thinned canopy locations were on average larger than their counterparts in intact-canopy sites, and exhibited greater iridescent blue and green coverage whereas fish from 'intact sites' had increased black coverage. Furthermore, Reznick *et al*. (2019) observed greater population density fluctuations in the thinned-canopy sites, where they reached much higher peak densities during the dry season compared to intact sites. The low densities in the rainy season, however, were generally similar for all four populations, with the exception of IT in the first year. Despite these phenotypic and population dynamics differences between the ecological treatments, we did not observe any effect on the genome-wide level. It is likely the phenotypes experienced strong plastic responses to the environmental changes, or that rapid phenotypic plasticity is followed by slower genetic assimilation (Price *et al.*, 2003; Torres-Dowdall, Handelsman, Reznick, *et al.*, 2012). It would also be interesting to identify selected alleles (the subject of Chapter 5) and see if/how they are affected by different environmental regimes.

Several population samples in our study had a strongly skewed sex ratio. ILL had three times as many females compared to males, whereas IUL and IC had nearly three times as many males. This could potentially affect genetic estimates if strong sexual selection plays a role in these populations. Dubois *et al.* (2018) showed that a skewed sex ratio did not affect downstream population genetics analyses. We found that most windows overlapped within an introduction year (e.g. between IUL and ILL and between IC and IT), despite the opposite sex skew between ILL and IUL. This suggests that the overlapping outlier windows we find

are not driven by the sex ratio. In addition, when we investigated per-sex nucleotide diversity, we found that male and female nucleotide diversity were not significantly different for ILL (skewed sex ratio, p=0.1655) and IT (equal sex ratio, p=0.08606). IUL, IC, GHP and GLP did have significantly different male/female nucleotide diversities (p<0.0001 for IUL, IC and GLP, p=0.01236), but the differences were small for IUL, IC and GHP, with female nucleotide diversity being slightly lower than male nucleotide diversity (7.0%, 8.2% and 2.1% lower respectively). These results show no consistent pattern of skew in nucleotide diversity, and therefore it is unlikely the skewed sex-ratios have affected our analyses.

Our findings of limited loss of genetic diversity in the introduced populations show that introducing populations to a new environment does not necessarily result in genetic bottlenecks and/or severe loss of genetic diversity, as is often assumed (Bodkin *et al.*, 1999; Mock *et al.*, 2004). All four populations were succesful and seem to be thriving. Based on our findings, we thus show the potential of using translocations to sucessfully (re-)introduce a species to a new environment, a finding that could help develop future conservation efforts. In addition we were unable to unravel the individual demographic histories of the introduced populations, most likely due to the lack of genetic differentiation after just a few generations. Elleouet and Aitken (2018) reported similar problems for ABC methods when inferring parameters of recent demographic events, and Benazzo *et al.* (2015) struggled to obtain parameter estimates for populations with low levels of genetic differentiation using both (an earlier version of) FASTSIMCOAL2 and ABC methods.

We have shown here coalescent modelling using the SFS could not resolve the demographic histories of these very young populations. We were not able to

identify a single best model for three of the four populations, and in the case of ILL our simulated parameter distributions were unrealistic and did not match observed data. It is likely that the very short time since the introduction has not allowed for enough population divergence in allele frequency to occur (either through genetic drift or new mutations), and therefore the populations reflect their ancestor's (GHP) demographic history rather than their own (Depaulis *et al.*, 2003; Sard *et al.*, 2019). If any new mutations have arisen during the 8 to 10 generations since the introduction, they would only be present at low frequencies, and in order to detect these rare variants we would need much larger sample sizes (N>100) (Beichman *et al.*, 2018). Methods based on the SFS (such as FASTSIMCOAL2) do not take linkage between SNPs into account (Excoffier *et al.*, 2013; Beichman *et al.*, 2018), even though this could provide important information about very recent events. Therefore, exploring methods that can include this information, such as the approximate Bayesian computation (ABC) approach (Beaumont *et al.*, 2002), or using segments of identity-by-descent (IBD) and identity-by-state (IBS) (Harris and Nielsen, 2013; Browning and Browning, 2015), could be useful to infer very recent demographic histories. In conclusion, we recommend exploring other methods, such as those mentioned above, instead of coalescent SFS-based methods for the inference of very recent demographic events, as we have shown here the coalescent was not able to resolve the population histories.

Overall, we found only minor genome-wide genetic differences between the four introduced populations and the HP source, and limited evidence for founding bottlenecks in three of the four introduced populations. Since the populations have only been diverging for a very short time, and the onset of selection on male size and age at maturity was established to be just two years ago (Reznick,

Bassar, *et al.*, 2019), it is very likely no new beneficial mutations have occurred yet, and allele frequencies in the introduced populations will still largely resemble those of the GHP source. We therefore predict that standard measures for detecting selection, such as F_{ST} outlier scans, might not be sensitive enough to detect these small shifts in allele frequencies (Stephan, 2016; Berner, 2019). In the next chapter, we will use the information about the limited differences between GHP and the introduced populations when looking for signals of selection along the genome.

Chapter 5 – DETECTING PARALLEL SELECTION IN RAPIDLY EVOLVING EXPERIMENTAL POPULATIONS OF GUPPIES

5.1 Abstract

Phenotypic evolution can occur more quickly than previously thought but understanding the genetic basis of rapid evolution is still in its infancy. Until recently it was difficult to detect such rapid adaptation at the genomic level. A unique opportunity to investigate rapid genomic adaptation is provided by four replicated experimental populations of guppies in the Northern Range Mountains of Trinidad that were transplanted from a high-predation (HP) source to guppy-free low-predation (LP) sites. Here, we compared whole genome variation in these recently introduced populations and their HP source. We uncovered signatures of selection with a combination of genome scans and a novel multivariate approach based on allele frequency change vectors. We were able to identify a limited number of candidate loci for convergent evolution across the genome. In particular, we found a region on chromosome 15 under strong selection in three of the four populations, with our multivariate approach revealing subtle parallel changes in allele frequency in all four populations across this region. Investigating patterns of genome-wide selection in this uniquely replicated experiment offers remarkable insight into the mechanisms underlying rapid adaptation, providing a basis for comparison with other species and populations experiencing rapidly changing environments.

5.2 Introduction

Over the years it has become clear that evolutionary change can be so rapid it occurs on an ecological timescale, and many studies have now shown that phenotypic traits can evolve within a few generations. Some examples include beak size changes in Darwin's finches (Grant and Grant, 2008), limb morphology of *Anolis* lizards (Losos, 2009), and colour variation in guppies (Endler, 1980). It remains largely unclear, however, what the genetic architecture underlying these rapidly changing phenotypes is.

Until recently it has been difficult to detect recent genomic adaptation. Firstly, rapid genetic adaptation is likely to occur through soft sweeps on standing genetic variation (SGV) because the beneficial alleles are already present in the populations, compared to hard sweeps where it takes time for beneficial *de novo* mutations to appear (Hermisson and Pennings, 2005; Barrett and Schluter, 2008). Because soft sweeps have weaker effects on linked sites, they are more difficult to detect than hard sweeps (Pritchard *et al.*, 2010). Secondly, whole genome sequencing (WGS) for non-model organisms is rare, and reduced representation approaches (such as microsatellites and RAD-seq) do not provide the level of detail required to detect the majority of selective sweeps (Tiffin and Ross-Ibarra, 2014; Lowry *et al.*, 2017). Now, with the advent of WGS for non-model organisms giving us access to genomes with unprecedented resolution, and improved methods for detecting soft sweeps, we can begin to answer this long-standing question in evolutionary biology.

In a soft sweep, multiple haplotypes containing the beneficial allele are increasing in frequency simultaneously, leaving behind a more subtle signature of reduced diversity and smaller shifts in allele frequency spectra (Hermisson and Pennings, 2005; Barrett and Schluter, 2008). In cases of recent adaptation

however, linkage disequilibrium (LD) will not yet have degraded, leaving a trace of haplotype homozygosity that can be detected with the extended haplotype homozygosity (EHH) metric (Sabeti *et al.*, 2002). Several methods have now been based on EHH, including a cross population measure (XP-EHH) and one specifically developed to detect soft sweeps (iHH12) (Garud *et al.*, 2015). iHH12 collapses the two most common haplotypes into one, making it more effective at detecting soft sweeps.

Finally, the availability of WGS for non-model organisms also provides an opportunity to look for more subtle shifts in allele frequency across many loci simultaneously, a signal of polygenic selection (Pritchard and Di Rienzo, 2010). Studies from the quantitative genetics field suggest that many traits under selection are polygenic (Falconer and Mackay, 1996; Barton and Keightley, 2002) and phenotypic evolution could occur very rapidly via minor allele frequency changes at many loci (Chevin and Hospital, 2008; Jain and Stephan, 2015). Considering the frequency shifts of alleles simultaneously at all loci involved (instead of individual SNPs) might be the best approach for detecting polygenic selection (Jain and Stephan, 2017).

In rapid adaptation scenarios, genomic changes are often small and difficult to distinguish from changes that occurred through random genetic drift. Populations that have evolved convergent phenotypes in response to similar environmental changes provide an opportunity to leverage signatures of repeated selection to detect subtle changes. The occurrence of convergent evolution is considered strong evidence for natural selection, as processes other than selection (such as genetic drift and random mutations) are unlikely to result in the same evolutionary changes in independent populations. Here, following Arendt and Reznick (2008), we use the term convergence to define the repeated evolution of similar traits

across independent populations, and make no distinction between convergent and parallel evolution. By looking for signatures of convergent evolution in rapidly evolving, replicated populations we will be able to distinguish random genomic changes from those caused by selection.

A unique opportunity to investigate rapid genomic adaptation is provided by four replicated experimental populations of guppies in the Northern Range Mountains of Trinidad. A more detailed description of this experiment can be found in Chapter 4 and Arendt *et al*. (2014). Briefly, in 2008 and 2009, juvenile guppies were collected from a single high predation (HP) site in the Guanapo river (GHP) and transplanted into four previously guppy-free LP locations; Lower Lalaja (ILL) and Upper Lalaja (ULL) in 2008 and Caigual (IC) and Taylor (IT) in 2009. ILL and IUL were introduced with lower initial densities than IC and IT due to less introduced fish and larger stream size than IC and IT, and each introduction year had one site where the canopy was thinned (IUL and IT) and one site with unchanged canopy density (ILL and IC). Kemp *et al*. (2018) and Reznick *et al*. (2019) found that these introduced guppies evolved convergent phenotypes similar to those observed in natural low predation (LP) populations: introduced males evolved increased colouration and were older and larger at the age of maturity. Additionally, population response varied by ecology; male guppies in thinned canopy sites evolved to have more iridescent green/blue (Kemp *et al.*, 2018) and experienced greater seasonal fluctuations in population densities than the closed-canopy populations (Reznick, Bassar, *et al.*, 2019). Reznick *et al*. (2019) also found that the onset of evolution for size and age at maturity differed between the density treatments, where the low density populations took a year longer to show evidence of selection. Finally, in Chapter 4 I showed that as a result of these differences in initial densities, ILL and IUL

have higher neutral diversity, likely because they were able to rapidly expand for a longer period before reaching peak densities.

Here we capitalise on this unique set of rapidly evolving experimental populations to investigate the early stages of adaptation. By using a combination of genome scans and multivariate analysis of allele frequency changes we identify signals of selection across the genome. The replicated nature of this experiment will allow us to distinguish signals of selection from those resulting from genetic drift by comparing the signals among the populations.

5.3 Methods

Genomic data set

For the analyses in this chapter, we use the same genomic data set of the recent introduction experiment that was described in Chapter 4. Briefly, approximately 20 fish were sampled from each location in the spring of 2013 (N=94). DNA was extracted using Qiagen DNeasy Blood and Tissue kit, and whole genome sequences were obtained with Illumina HiSeq technology. GATK's Best Practices protocol was used to process the raw data (v3.8 & 4.0) (van der Auwera *et al.*, 2014). A more detailed description of this process can be found in Chapter 4. The final VCF file contained 5,700,890 SNPs. For analyses based on haplotypes, VCF files were phased as in Malinsky *et al*. (2018), using BEAGLE (Browning and Browning, 2007) and SHAPEIT2 (Delaneau *et al*., 2012) for optimum phasing quality.

Outlier scans

To identify signals of selection along the genome, we used an outlier window approach. The introduced populations have only recently been established and

we found only slight differences in the neutral evolution and population differentiation between these populations and the GHP source population (see Chapter 4). To gain insight into the allele frequencies of the populations, we first investigated how many fixed differences exist between GHP and each LP population, as well as windowed allele frequency changes (∆AF) between GHP and each of the LP populations by calculating the change in minor allele frequency in non-overlapping sliding windows of 75000 bp as follows:

$$|\Delta AF| = abs(AF_{LP} - AF_{HP})$$

Informed by those results, we chose to use two statistics based on haplotype homozygosity which are more suitable for detecting very recent directional selection as recombination will not have broken up the selected haplotypes. The first measure is cross-population extended haplotype homozygosity (XP-EHH), which was developed to detect selective sweeps that are (nearly) fixed in one populations, but still polymorphic in the population as a whole (Sabeti *et al.*, 2007). XP-EHH compares haplotype lengths across two populations to control for local variation in recombination rates, and by normalising XP-EHH values it corrects for genome-wide differences in haplotype length between populations. Therefore, we also used iHH12 (Garud *et al.*, 2015; Torres *et al.*, 2018), which combines the top two most frequent haplotypes into a single haplotype, making it more effective at detecting soft sweeps. Both XP-EHH and iHH12 were calculated using SELSCAN v1.2.0a (Szpiech and Hernandez, 2014), using the default settings, except we applied a MAF filter of 0.01. The results were then frequency-normalised over all chromosomes using the script provided by SELSCAN. XP-EHH outliers were identified by a value of XP-EHH > 2.5, and iHH12 outliers with an absolute value of iHH12 > 5.

The number of overlapping outliers among populations per measure was calculated using the R package SuperExactTest (Wang *et al.*, 2015). This package also calculates the expected number of overlapping windows for each set based on a hypergeometric distribution.

Allele frequency changes in outlier windows

In young populations it is likely changes in allele frequency are minor. Therefore, to assess if we missed occurrences of parallelism that were not detected with the haplotype homozygosity methods, we investigated whether similar allele frequency changes (∆AF) can be observed in the XP-EHH and iHH12 outlier windows in all four introduction populations. For this analysis, we narrowed down the number of outlier windows by only including regions with a minimum of three consecutive outlier windows and plotted these onto pair-wise allele frequency changes plots. If outlier SNPs are evolving in parallel among populations, we expect them to fall on the diagonal line (x=y), which indicates ∆AF of the same size and in the same direction. Therefore, if SNPs ∆AF is off the diagonal, one population experienced a bigger change in allele frequency than the other, or a change in a different direction, indicating non-parallel evolution. If a SNP falls on the perpendicular, it suggests they experienced similar sized changes, but in opposite directions, which could indicate anti-parallel evolution (figure 5.1).

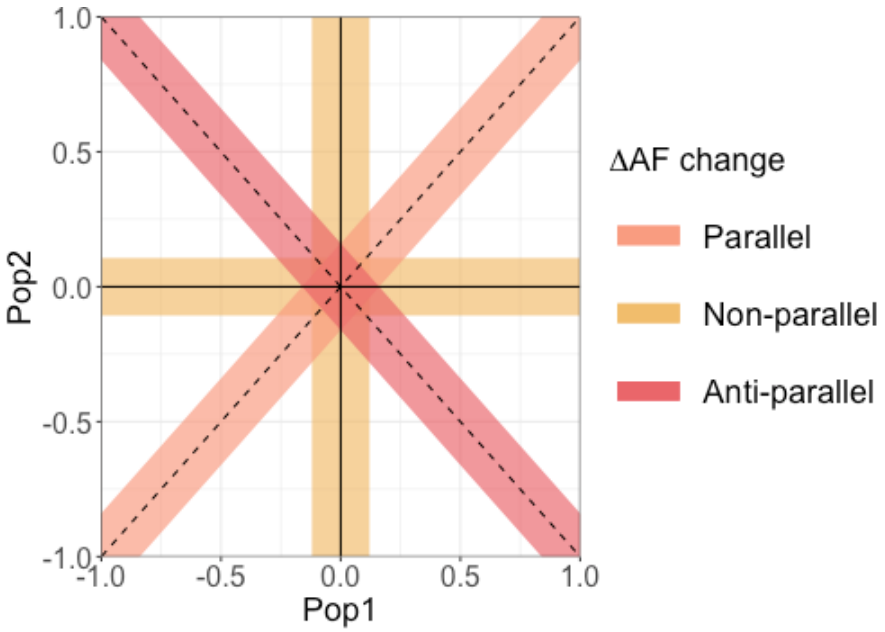


Figure 5.1 Schematic showing parallel, non-parallel and anti-parallel changes in allele frequency between two populations.

Multivariate analysis of parallel allele frequency changes

To investigate parallel allele frequency changes in all four populations simultaneously, we adapted De Lisle and Bolnick's (De Lisle and Bolnick, 2020) multivariate approach for interpreting parallel evolutionary change. Instead of using phenotypic change vectors, we applied the method to vectors of allele frequency change between the HP source and each of the introduced populations. Briefly, we divided each chromosome into non-overlapping windows of 500 SNPs, and calculated a matrix of normalised allele frequency changes per window:

$$X = \begin{matrix} \Delta AF_{1,1} & \dots & \Delta AF_{n,1} \\ \vdots & \ddots & \vdots \\ \Delta AF_{1,m} & \dots & \Delta AF_{n,m} \end{matrix}$$

Where *n* is the number of SNPs in a window and *m* the number of populations. For each of these matrices we then computed the correlation matrix describing

correlations of normalised allele frequencies among the vectors of each HP-LP pair, by multiplying matrix *X* with its transpose:

$$C=XX^T$$

Eigen decomposition of each C matrix produces a distribution of eigenvalues, where distributions with excessive variance explained by the first eigenvector suggest a shared direction of allele frequency changes through multivariate space for all population pairs. This direction of change could be parallel (along the same axis in the same direction) or anti-parallel (along the same axis in different directions). In order to determine which windows show extreme eigenvalues, we created a null-distribution. We randomly sampled windows of an equal number of SNPs (N=500) along the genome, but allowed for window start positions to not match those of the observed dataset. For each null permutation, individual IDs were shuffled and allele frequency vectors re-calculated between random pairs. To each window, we applied the same transformations described above and permuted it 1000 times. Windows were identified as outliers if they had an eigenvalue above the genome-wide 99.9% quantile of the null-distribution. Finally, by examining the loading of each population on the first eigenvector, we investigated where the allele frequency changes are parallel and where they are anti-parallel. Lineages with the same sign loading can be interpreted as evolving in the same direction, and lineages with opposite signs are evolving anti-parallel.

To identify candidate regions of selection, we identified the overlapping regions between the outlier regions on eigenvector 1 and the XP-EHH and iHH12 outlier windows. We also investigated linkage disequilibrium (LD) across the candidate regions. For this analysis, we extracted the SNPs located in the overlapping outlier regions from the main VCF file and using PLINK *--r2 square* (v1.90b (Chang

et al., 2015)) calculated LD between all pairs of SNPs. In total, there were 6204 SNPs in the outlier region, resulting in 38,489,616 pairs of LD calculations.

5.4 Results

In agreement with our findings in Chapter 4, where we only found minor changes in neutral diversities, we found little change in allele frequency in the introduced populations. Mean allele frequency change was low in all populations (table 5.1, figure 5.2). There were no fixed differences in allele frequency between GHP and any of the four introduction populations, whereas we identified 1,770 fixed differences between GHP and GLP (table 5.1). ILL and IUL had fewer SNPs that were the minor allele in GHP and became fixed in the introduced sites compared to IC and IT (N=120 and N=132 in ILL and IUL, and N=343 and N=596 for IC and IT, respectively, table 5.1). None of these fixed SNPs were shared among all four populations. IC and IT had more sites where allele frequencies had not changed from those in GHP compared to ILL and IUL (table 5.1), consistent with the shorter period of random drift they experienced before the onset of selection in these populations (Reznick, Bassar, *et al.*, 2019). Finally, GLP had several orders of magnitude more fixed SNPS than the introduced populations (N=480,372, table 5.1).

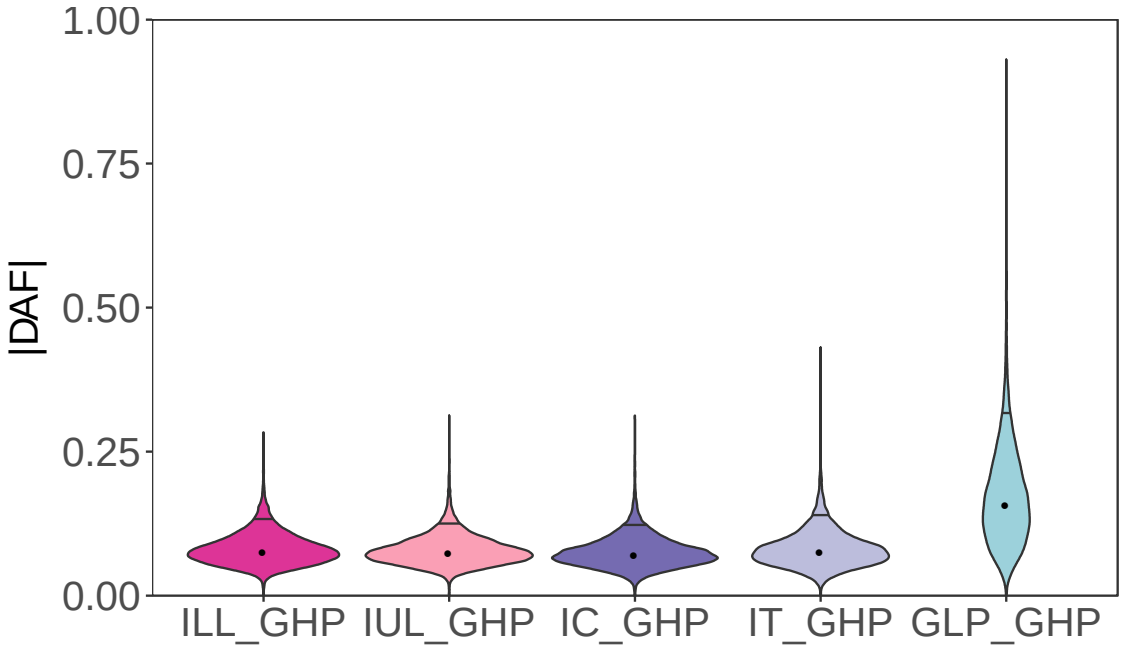


Figure 5.2 Distributions of minor allele frequency changes between GHP and each of the LP populations. Black points indicate median, black line the 95% confidence interval.

Table 5.1 Summary of allele frequency changes in all populations. Fixed GHP-LP shows the fixed differences between GHP and each of the LP populations. GHP MF/fixed LP shows the number of SNPs that were present as the minor allele in GHP, but are fixed in the LP populations. Unchanged MAF shows how many SNPs had the same MAF for both GHP and LP.

Population	ID	Fixed GHP-LP	GHP MAF/ fixed LP	Unchanged MAF	mean |ΔAF|	median |ΔAF|	Min-max
Lower Lalaja	ILL	0	120	478833	0.082	0.058	0-0.719
Upper Lalaja	IUL	0	132	506220	0.080	0.056	0-0.753
Caigual	IC	0	343	893384	0.076	0.053	0-0.802
Taylor	IT	0	596	893015	0.082	0.056	0-0.843
Guanapo LP	GLP	1770	480372	1102865	0.173	0.083	0-1.000

Outlier window analysis

The marginal differences found above and in Chapter 4 suggest allele frequency based genome scans such as F_{ST} might not be suitable here, as they would not be able to detect minor changes in allele frequencies (Berner, 2019). We therefore chose to use two haplotype-based methods, XP-EHH and iHH12, for the outlier window analysis as they are better suited to detect very recent

selection. We analysed a total of 9,878 windows of 75kb length and used a cut-off of >2.5 for the normalised XP-EHH values and of >5 for iHH12 (figure 5.3). For the XP-EHH statistic, we found IT had the most outlier windows (N=57), followed by ILL (N=38), IC (N=22) and finally IUL (N=17) (figure 5.3a, supplementary tables S1-4). Similarly, for iHH12, IT had the most outliers (N=40), followed by IC (N=25), ILL (N=23) and IUL (N=15) (figure 5.3b, supplementary tables S5-8).

By comparing outlier windows among populations, we found that most XP-EHH outliers were unique. There were no windows that overlapped among all four populations for XP-EHH (figure 5.4a). The three-way comparison among ILL, IC and IT had the highest fold enrichment of all comparisons and contained three overlapping outlier windows (expected N = <1). These overlapping outliers were all located on chromosome 15 (table 5.2). Comparisons with IUL had the fewest overlapping outlier windows in pair-wise comparisons: one each for the pairwise comparisons with ILL, IT and IC. Among the pairwise comparisons, ILL-IT overlapped most (N=6), followed by IC-IT (N=5) and ILL-IC (N=4) (figure 5.4a).

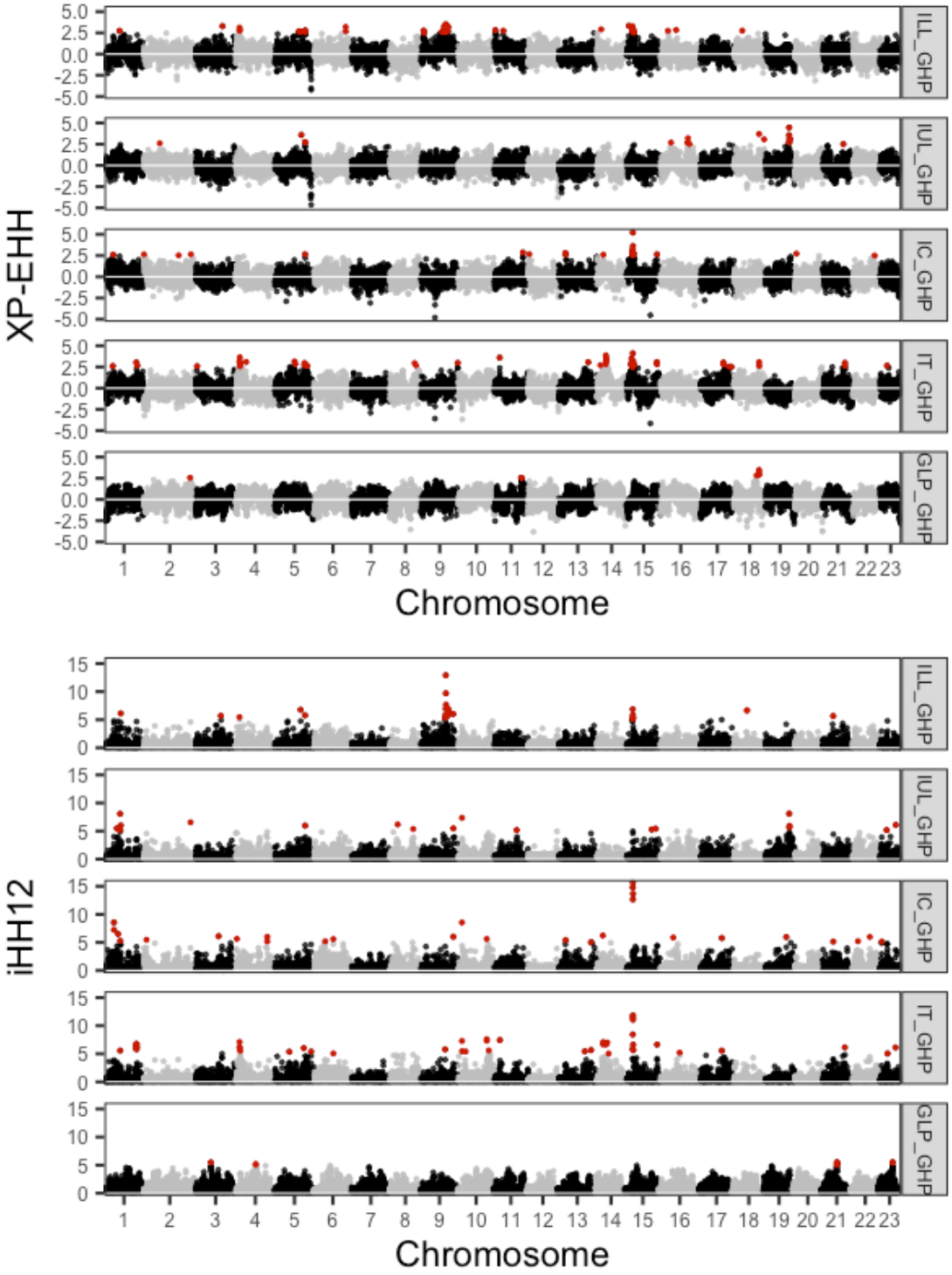


Figure 5.3 Genome-wide values of haplotype genome scans. A) XP-EHH and B) iHH12 for each LP population. Red points mark outlier windows, where outliers are defined as XP-EHH > 2.5 and iHH12 > 5.

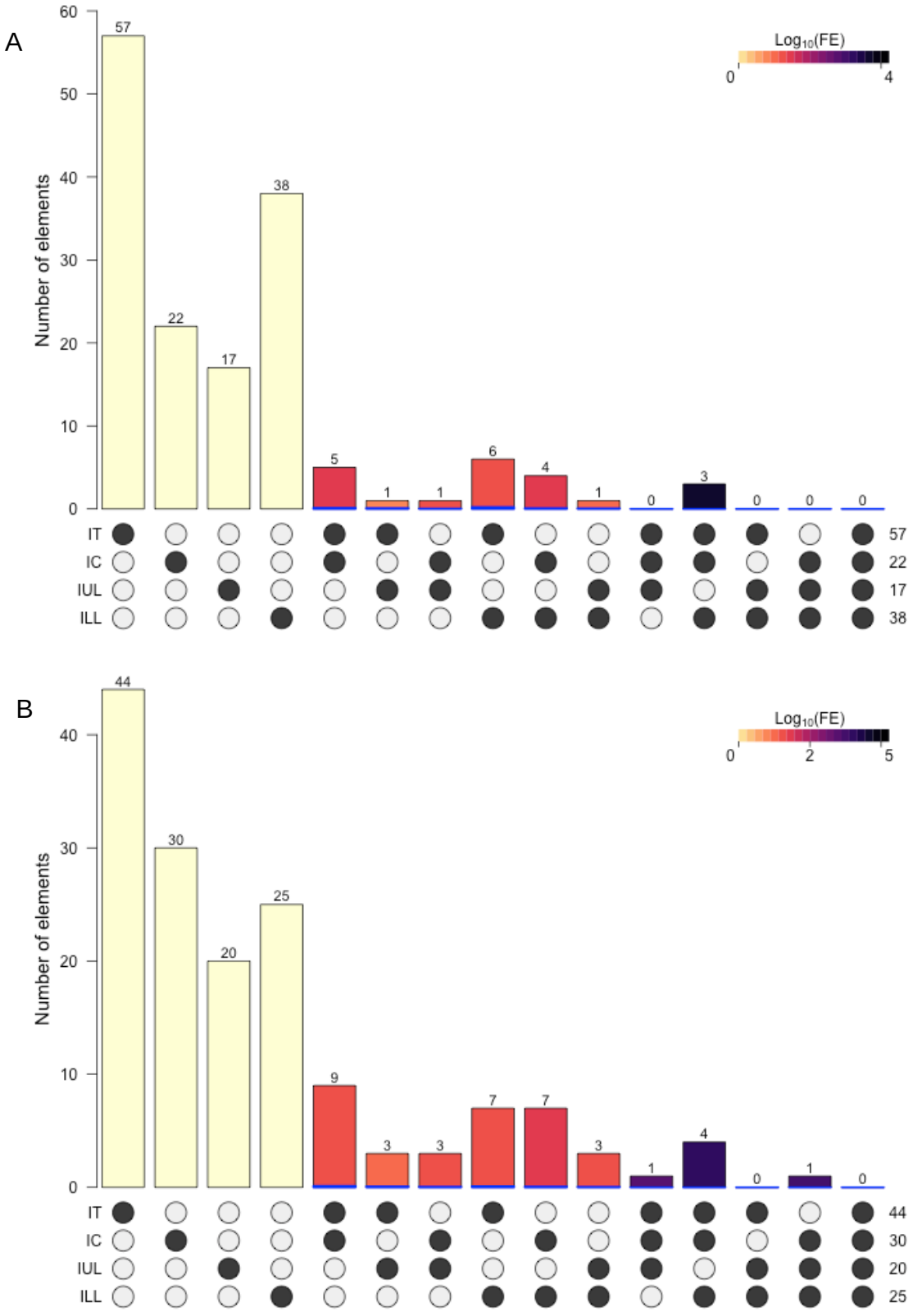


Figure 5.4 Overlap of genome scan outliers among the four introduced populations for A) XP-EHH and B) iHH12. Blue boxes represent the expected number of overlapping outlier windows. Colour scale indicates log-scaled fold enrichment.

Similarly, iHH12 outliers were found to have little overlap among the populations. Again, no outlier windows were found in all four populations and ILL, IC, and IT shared four overlapping windows (expected N = <1, figure 5.4b), all located on chromosome 15 (table 5.2). These windows on chromosome 15 are the same as the three outlier windows found with XP-EHH for these three populations. We found one overlapping window on chromosome 9 (27,750,001-27,825,000bp) among ILL, IUL and IC, and one overlapping window on chromosome 10 (2,850,001-2,925,000bp) among IUL, IC and IT. Similar results as those observed for XP-EHH were found for pair-wise comparisons, with comparisons including IUL showing fewer overlapping outlier windows. Each comparison with IUL had three overlapping windows. Among the pairwise comparisons, IC-IT had the most overlapping windows (N=9), followed by ILL-IT and ILL-IC (N=7 for both). Taken together, these results suggest a region on chromosome 15 is a strong candidate for rapid adaptation in the introduced populations.

Table 5.2 Overlapping outlier windows among all populations for XP-EHH and iHH12.

Comparison	Overlap XP-EHH	Overlap iHH12
ILL & IUL	chr5:26175001-26250000	chr1:14250001-14325000 chr5:26025001-26100000 chr9:27750001-27825000
ILL & IC	chr15:5025001-5100000 chr15:5400001-5475000 chr15:5475001-5550000 chr15:6225001-6300000	000104F_0:1200001-1275000 chr15:5175001-5250000 chr15:5250001-5325000 chr15:5325001-5400000 chr15:5400001-5475000 chr21:9000001-9075000 chr9:27750001-27825000
ILL & IT	chr4:3600001-3675000 chr4:3675001-3750000 chr15:5025001-5100000 chr15:5400001-5475000 chr15:5475001-5550000 chr15:6600001-6675000	chr4:3450001-3525000 chr15:5175001-5250000 chr15:5250001-5325000 chr15:5325001-5400000 chr15:5400001-5475000 chr15:5475001-5550000 chr9:20625001-20700000
IUL & IC	chr5:26100001-26175000	chr1:13800001-13875000 chr9:27750001-27825000 chr10:2850001-2925000
IUL & IT	chr18:21825001-21900000	chr1:13725001-13800000 chr10:2850001-2925000 chr23:13575001-13650000
IC & IT	chr15:5025001-5100000 chr15:5325001-5400000 chr15:5400001-5475000 chr15:5475001-5550000 chr15:26400001-26475000	chr6:16500001-16575000 chr10:2850001-2925000 chr10:24375001-24450000 chr13:27750001-27825000 chr15:5175001-5250000 chr15:5250001-5325000 chr15:5325001-5400000 chr15:5400001-5475000 chr17:18225001-18300000
ILL & IUL & IC	NA	chr9:27750001-27825000
ILL & IUL & IT	NA	NA
ILL & IC & IT	chr15:5025001-5100000 chr15:5400001-5475000 chr15:5475001-5550000	chr15:5175001-5250000 chr15:5250001-5325000 chr15:5325001-5400000 chr15:5400001-5475000
IUL & IC & IT	NA	chr10:2850001-2925000
ILL & IUL & IC & IT	NA	NA

Allele frequency changes in outlier windows

To investigate whether the populations experienced similar changes in allele frequencies in the outlier regions, we examined allele frequency change per SNP in strong candidate regions for selection. Here, we examined regions with at least three consecutive outlier windows for either XP-EHH and iHH12. This resulted in 33 outlier windows for XP-EHH, collapsing into six consecutive regions on five chromosomes and 20 windows for iHH12, collapsing to four regions on four chromosomes (tables 5.3 and 5.4). Overall, outlier SNPs showed similar changes between populations than non-outlier SNPs (figures 5.5 and 5.6). Interestingly, the SNPs of outlier windows on chromosome 15 are broadly located along the diagonal (especially for the iHH12 regions, figure 5.6), even for pairwise comparisons with IUL, for which we found no outlier regions on chromosome 15. This suggests that all four populations are evolving along the same axis for these SNPs.

Table 5.3 Overlapping outlier windows collapsed into regions of at least three consecutive windows for XP-EHH.

Chromosome	Start	End	Number of windows
chr4	3525001	3900000	5
chr9	21225001	21600000	5
chr14	7725001	8100000	5
chr15	4125001	4800000	9
	5175001	5625000	6
chr19	20475001	20700000	3

Table 5.4 Overlapping outlier windows collapsed into regions of at least three consecutive windows for iHH12

Chromosome	Start	End	Number of windows
chr1	27750001	28200000	6
chr4	3450001	3675000	3
chr9	21150001	21525000	5
chr15	5100001	5550000	6

However, we also observed SNPs that diverge substantially from the diagonal, indicating the two populations experienced different changes in allele frequency. For example, SNPs within the chromosome 19 region showed differing changes among the populations. First, IUL and ILL experienced a change of allele frequencies in opposite directions (figure 5.5a), meaning IUL saw a reduction in the minor allele frequency (indicated by a negative ΔAF) for SNPs where ILL experienced in increase in the MAF (positive ΔAF), which could indicate anti-parallel evolution (figure 5.1). Whereas, values of IC for chromosome 19 fall more on the diagonal when compared to IUL, suggesting IC and IUL are evolving in parallel for these SNPs. Finally, ΔAF values of IT for chromosome 19 SNPs are

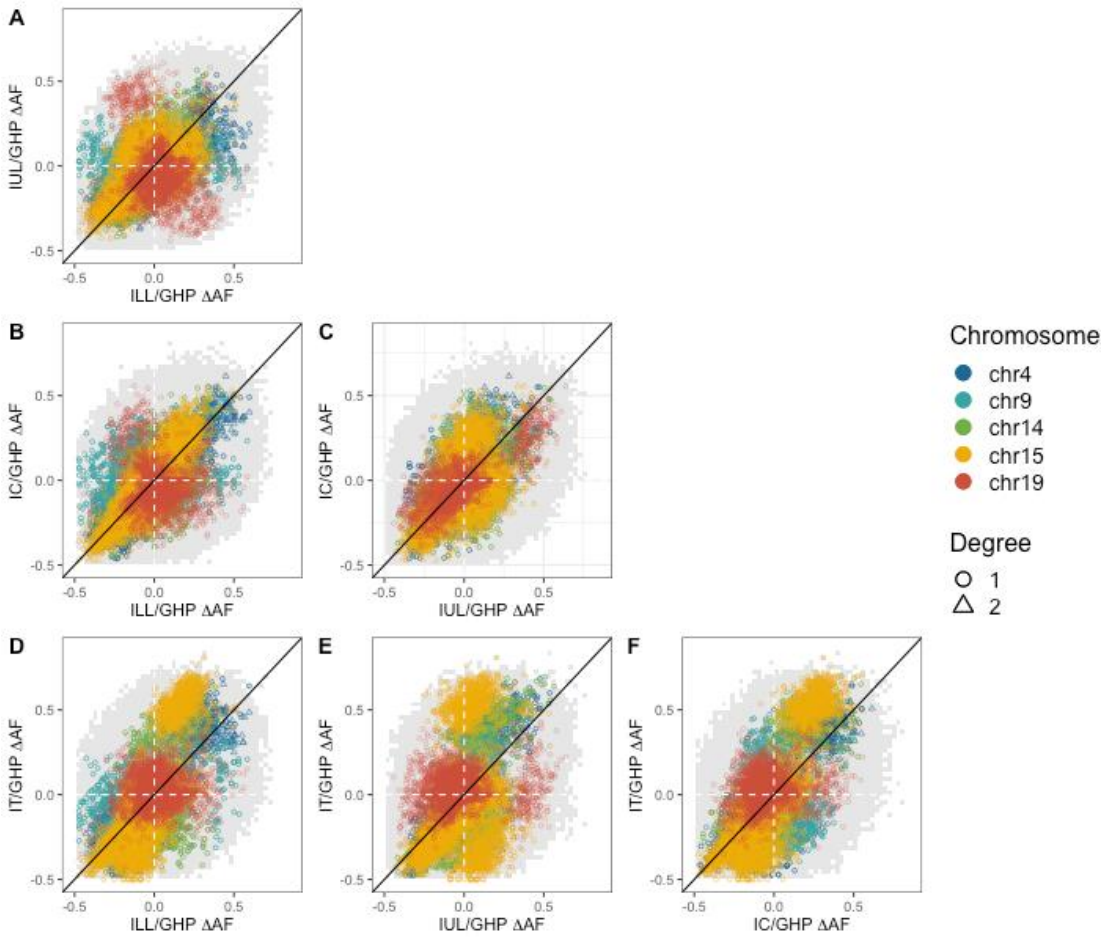


Figure 5.5 Allele frequency changes in XP-EHH outlier windows for all pairwise comparisons among the four introduction populations. The diagonal indicates the pair-wise comparisons had the same change in allele frequency. Grey indicates the allele frequency changes of non-outlier SNPs. Coloured spots indicate the SNPs in outlier windows, coloured by chromosome.

centred around zero, indicating no change in this region (figures 5d-f). Similarly, SNPs on chromosome 9 are divergent in pairwise comparisons with ILL, while values of ΔAF for IUL, IC and IT are centred around zero (figure 5.6 a, b, d) suggesting non-parallel evolution of these SNPs.

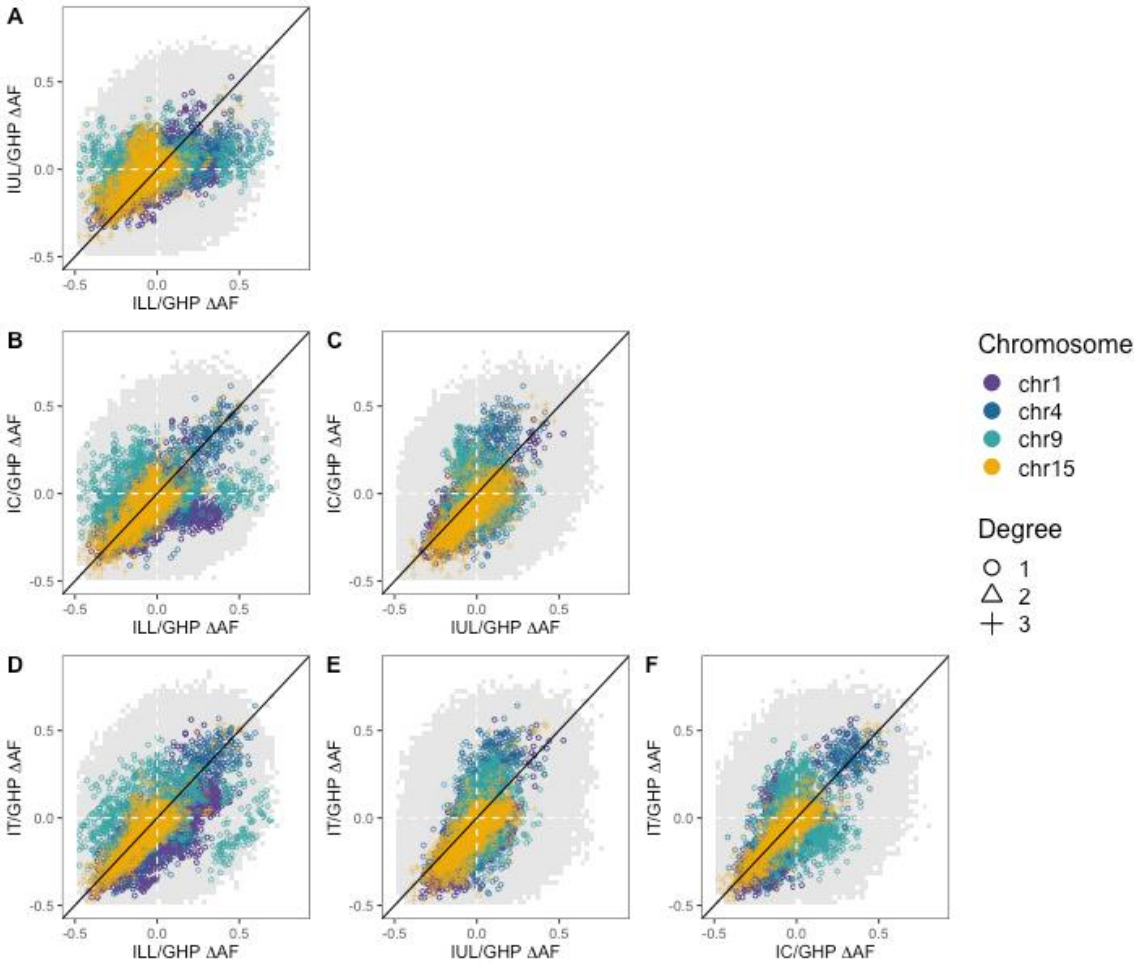


Figure 5.6 Allele frequency changes in iHH12 outlier windows for all pairwise comparisons among the four introduction populations. The diagonal indicates the pair-wise comparisons had the same change in allele frequency. Grey indicates the allele frequency changes of non-outlier SNPs. Coloured spots indicate the SNPs in outlier windows, coloured by chromosome.

Multivariate analysis of parallel allele frequency changes

We assessed whether the four populations exhibited parallel changes in allele frequency using a multivariate approach. This method considers changes in all populations simultaneously, and might therefore detect candidate regions where the populations experience parallel allele frequency changes that are too small to be detected by overlapping outlier methods. We identified 38 outlier windows on the first eigenvector, spread across ten chromosomes, using the 99.9% confidence interval cut-off from the randomised allele frequency matrices (figure 5.7a, supplementary table S5.9). The majority of the 38 windows were found on chromosome 8 (N=8), chromosome 10 (N=9) and chromosome 15 (N=8) (figure 5.8). In each outlier window, all populations had eigenvector loadings of the same sign (positive/negative) and similar size, suggesting all four populations experienced parallel allele frequency changes (along the same axis and in the same direction, supplementary table S5.9).

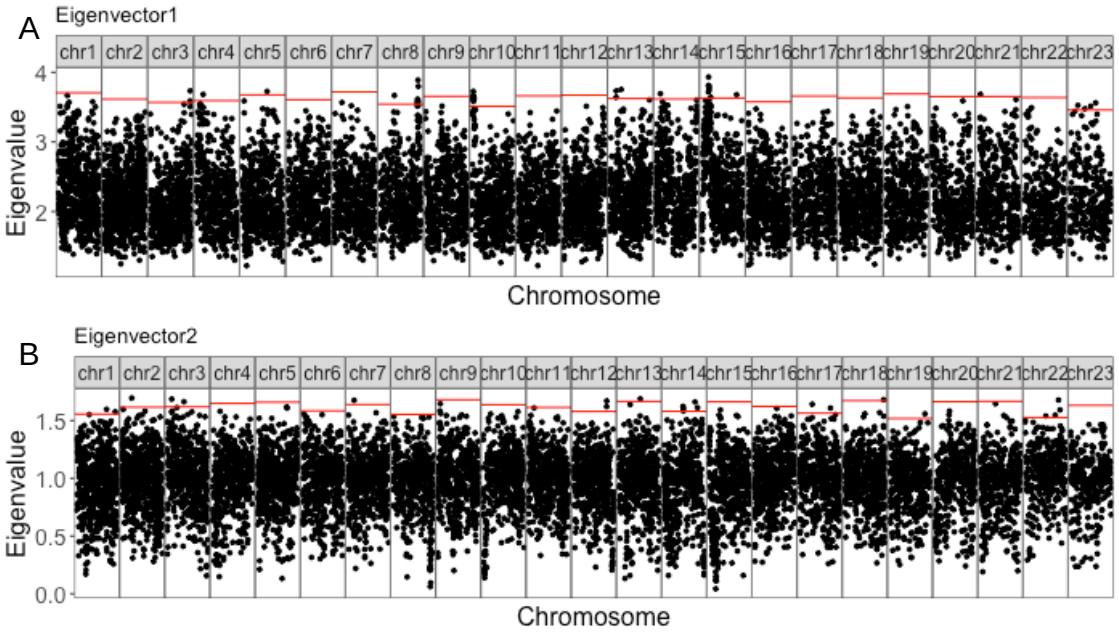


Figure 5.7 Eigenvalues for the first two eigenvectors along the genome. Red lines indicate the per-chromosome 99% confidence interval cut-off based on a randomised null distribution.

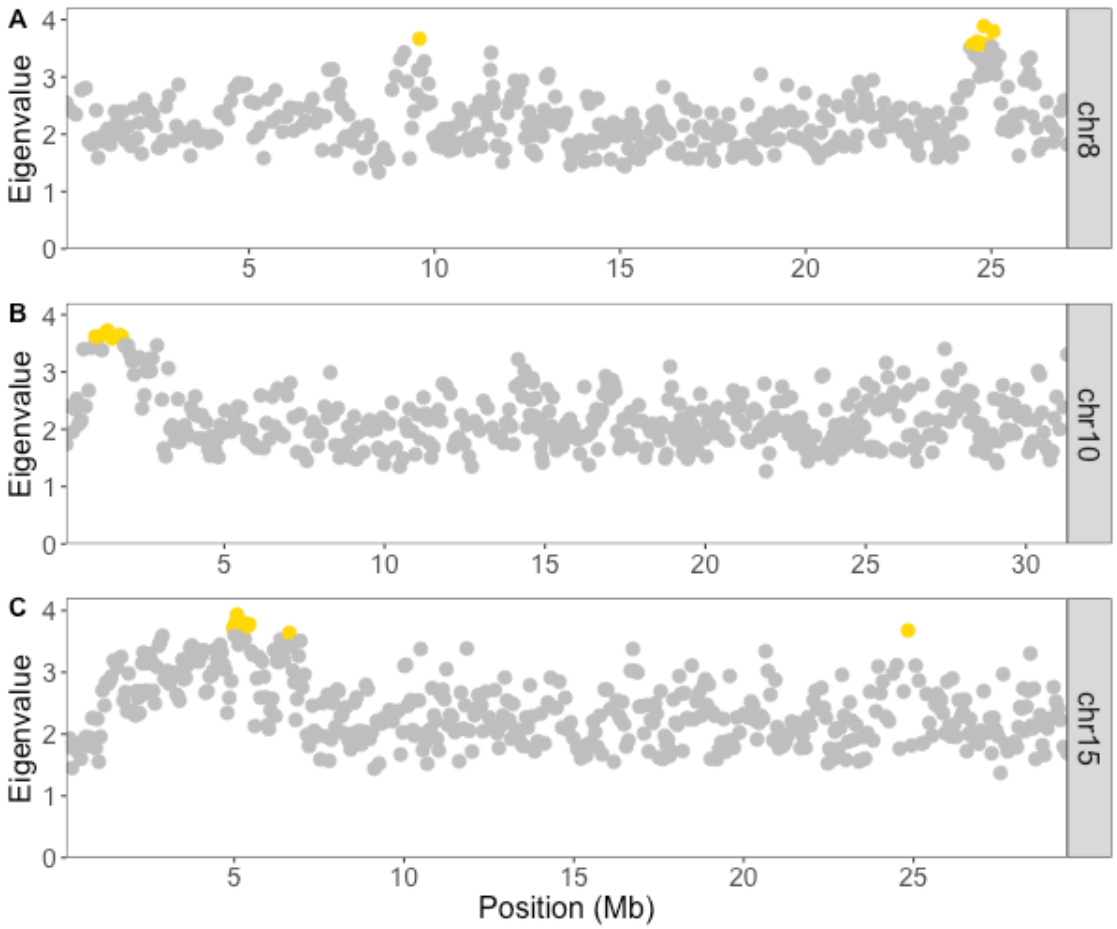


Figure 5.8 Eigenvalues for the first eigenvector along chromosomes with the most outlying windows. A) chromosome 8, B) chromosome 10 and C) chromosome 15. Yellow data points indicate windows with eigenvalues above the 99.9% confidence intervals.

Examining absolute allele frequency changes in the outlier windows revealed variation among the populations (figure 5.9). All populations had similar mean values of ΔAF although IT was slightly higher (IT = 0.151, ILL=0.139, IUL=0.139, IC=0.139). The consecutive outlier windows on chromosome 10 showed the largest AF change, although values in IC were consistently lower in this region. For the outlier window blocks on chromosome 8 and chromosome 15, IUL had consistently lower values of |ΔAF| than the other three populations (figure 5.9).

On eigenvector 2, we identified 18 outlier windows, spread across ten chromosomes (figure 5.7b, supplementary table S5.10). Most outlier windows were found on chromosome 14 (N=5). There were no windows that were outliers

on both eigenvector 1 and eigenvector 2, indicating only one dimension of shared evolutionary change.

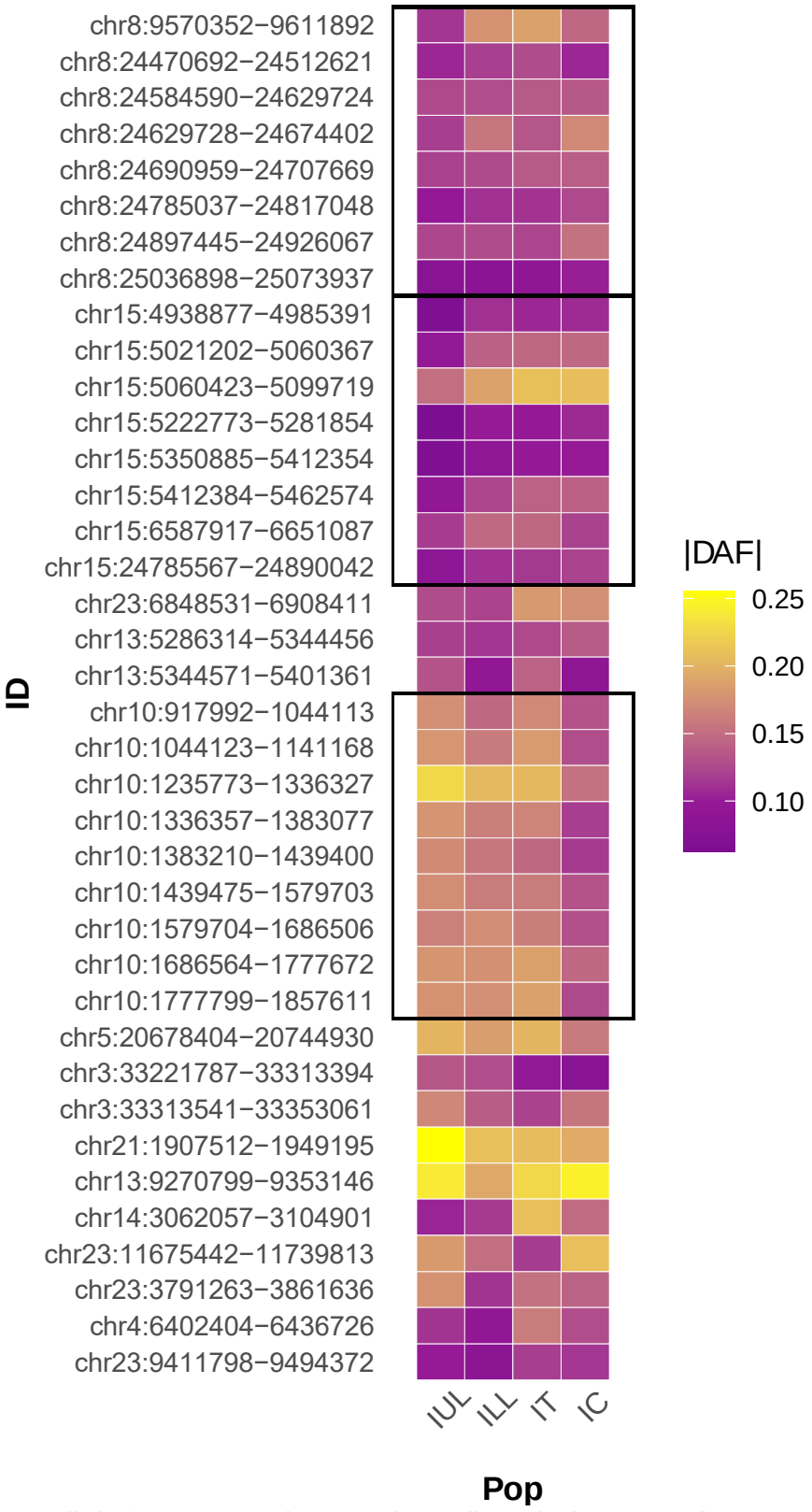


Figure 5.9 Allele frequency changes in outlier windows on eigenvector 1. Black boxes indicate the three largest consecutive outlier regions on chromosome 8, 10 and 15.

Candidate region on chromosome 15

Overlapping the outlier windows of each of our three methods (iHH12, XP-EHH and the eigenvector analysis) highlighted a region on chromosome 15 from 4,938,877 to 6,651,087 bp that contained seven outlier windows (figure 5.10, tables 5.5 & 5.6). One of these windows (5,060,423-5,099,719 bp) also had high values of |ΔAF| for all populations in the multivariate analysis (figure 5.9). Taking all the evidence together, this region on chromosome 15 is under strong selection in at least three of the four populations and experienced parallel allele frequency changes among all four populations. Taking a closer look at per SNP values of XP-EHH, iHH12 and ΔAF, we see that the signal is strongest in the region coding for two cadherin genes (figure 5.11). These genes have been found to be implicated in regulating pigment cell migration, and are involved in cell-cell adhesion interactions (Fukuzawa & Obika, 1995; Nishimura *et al*., 1999). A third gene identified in these windows is ras guanyl-releasing protein 3 (RASGRP3), which has been identified as a differentially expressed gene in a study investigating differential body size in mandarin fish (Tian *et al.*, 2016). Finally, looking at patterns of linkage disequilibrium in the entire region on chromosome 15, we found high LD across the region, especially in ILL, IC and IT (figure 5.12), a pattern consistent with an ongoing or recent selective event.

Table 5.5 Overlapping outlier windows between eigenvector 1 outliers and iHH12 outlier windows

	Eigenvector 1 outlier		**iHH12 outlier**		
Chromosome	**BP1**	**BP2**	**BP1**	**BP2**	**Genes**
chr15	5222773	5281854	5175001	5250000	ttc27
			5250001	5325000	
	5350885	5412354	5325001	5400000	ahnak, edaradd, pdzd8
	5412384	5462574	5400001	5475000	pdzd8, slc18a2, kcnk18

Table 5.6 Overlapping outlier windows between eigenvector 1 and XP-EHH outlier windows.

	Eigenvector 1 outlier		XP-EHH outliers		
Chromosome	**BP1**	**BP2**	**BP1**	**BP2**	**Genes**
chr5	20678404	20744930	20700001	20775000	rassf5, ikbke, srgap2
chr14	3062057	3104901	3075001	3150000	cd248, face2, trnae-cuc
chr15	4938877	4985391	4950001	5025000	LOC10347655, cdh1
	5021202	5060367	5025001	5100000	B-cadherin, cdh1
	5060423	5099719			B-cadherin, rasgrp3, ltbp1
	5222773	5281854	5175001	5250000	ttc27
			5250001	5325000	
	5350885	5412354	5325001	5400000	ahnak, edaradd, pdzd8
	5412384	5462574	5400001	5475000	pdzd8, slc18a2, kcnk18
	6587917	6651087	6600001	6675000	wdpcp

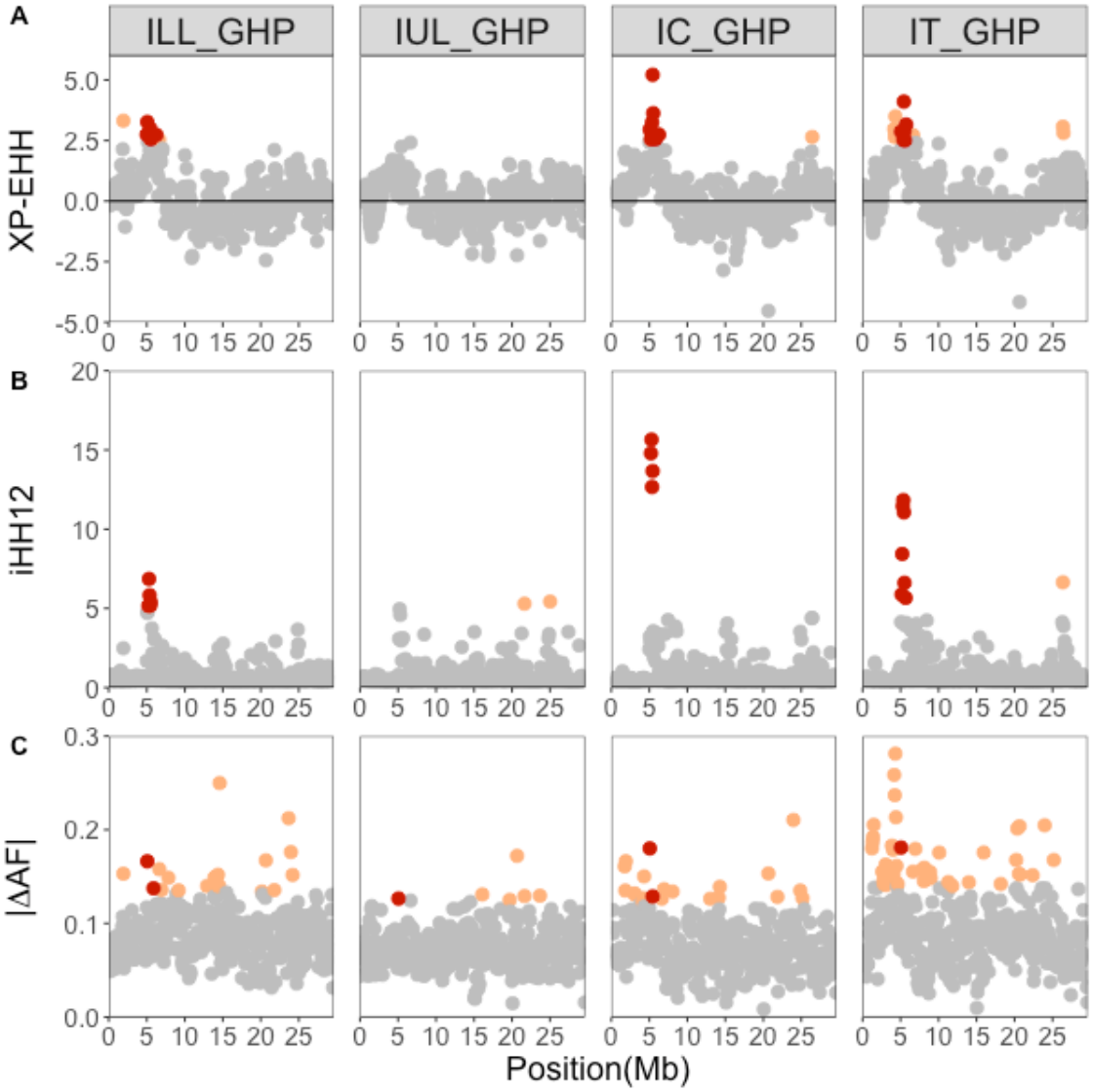


Figure 5.10 Per window values of XP-EHH, iHH12 and |ΔAF| on chromosome 15 highlighting the overlapping outlier windows between each measure and the outliers found on eigenvector 1 for each introduced population. A) XP-EHH, B) iHH12 and C) |ΔAF|. Orange dots mark outlier windows and red dots mark overlapping windows with XP-EHH > 2.5 or iHH12 > 5 and |ΔAF| > 95% confidence interval.

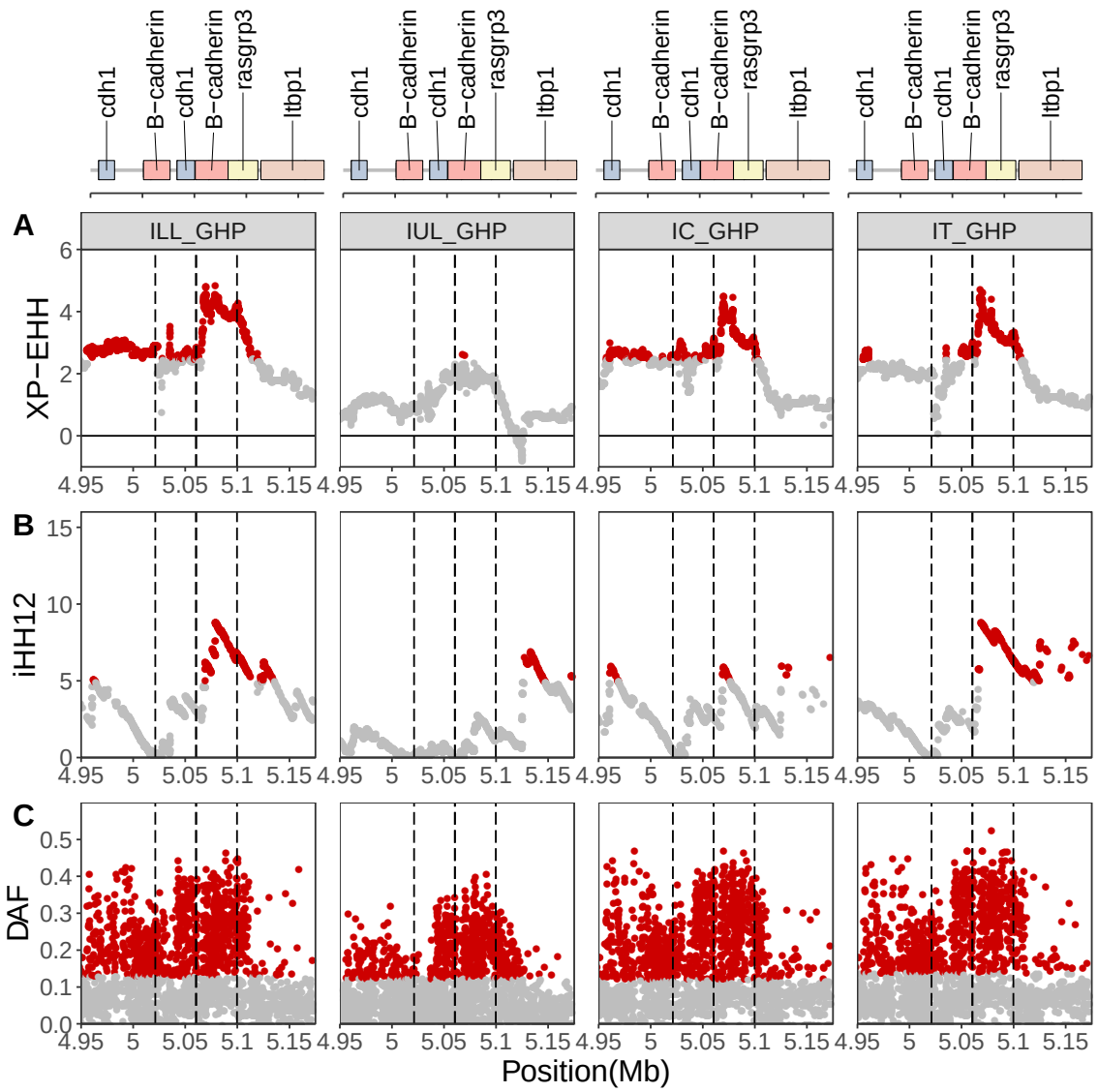
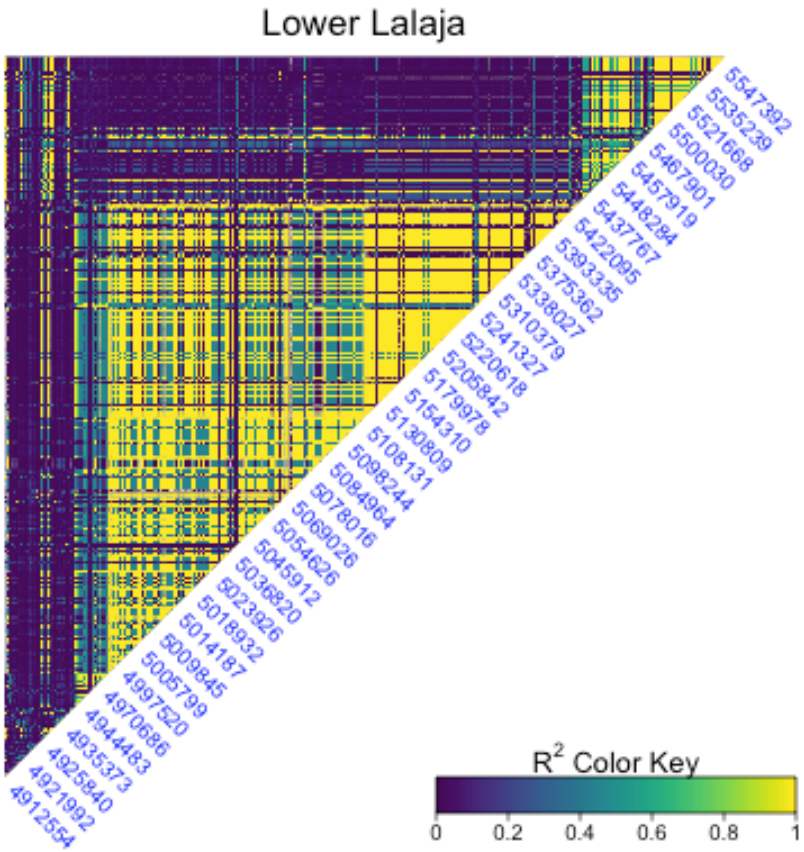
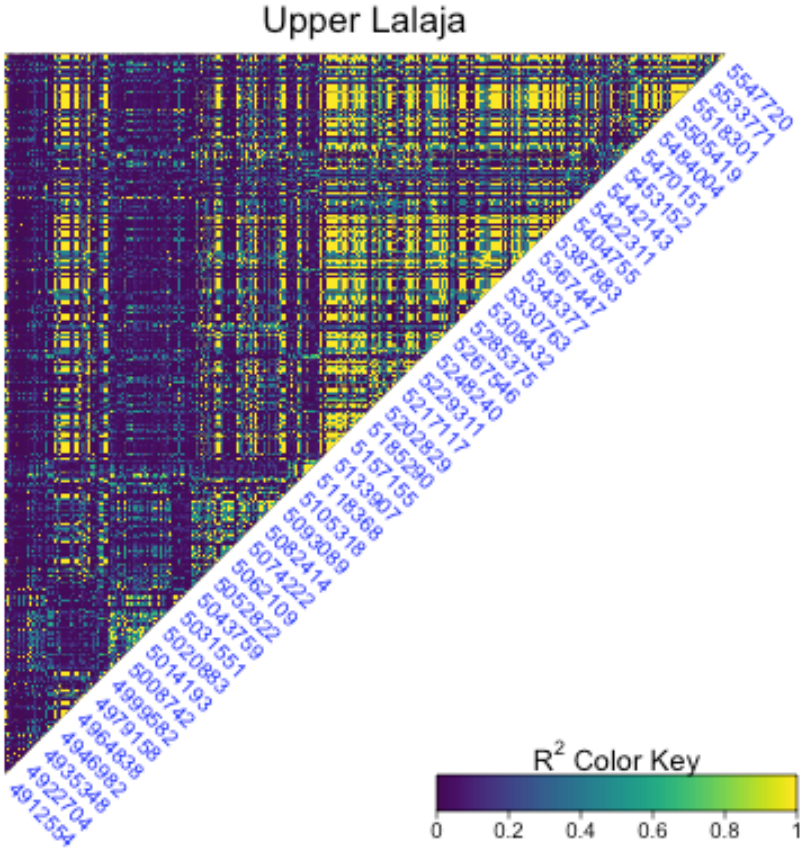


Figure 5.11 Per SNP values of XP-EHH, iHH12 and |ΔAF| for two consecutive windows on chromosome 15 between 5,021,202 bp and 5,099,719 bp. A) XP-EHH and B) |ΔAF|. Dashed lines indicate the location of the outlier windows in interest found with the eigenvector analysis, red dots mark outlier SNPs.

A


Lower Lalaja
5547392
5535239
5521668
5500030
5467901
5457919
5448284
5437767
5422095
5393335
5375362
5338027
5310379
5241327
5220618
5205842
5179978
5154310
5130809
5108131
5098244
5084964
5078016
5069026
5054626
5045912
5036820
5023926
5018932
5014187
5009845
5005799
4997520
4970686
4944483
4935373
4925840
4921992
4912554
R2 Color Key
0
0.2
0.4
0.6
0.8
1

B


Upper Lalaja
5547720
5533771
5518301
5505419
5494004
5470151
5453152
5442143
5422311
5404755
5387883
5367447
5343377
5330763
5308432
5285375
5267946
5248240
5229311
5217117
5202829
5185290
5157155
5133907
5118368
5105318
5093089
5082414
5074222
5062109
5052822
5043759
5031551
5020883
5014193
5008742
4996582
4979158
4964838
4946982
4935348
4922704
4912554
R2 Color Key
0
0.2
0.4
0.6
0.8
1

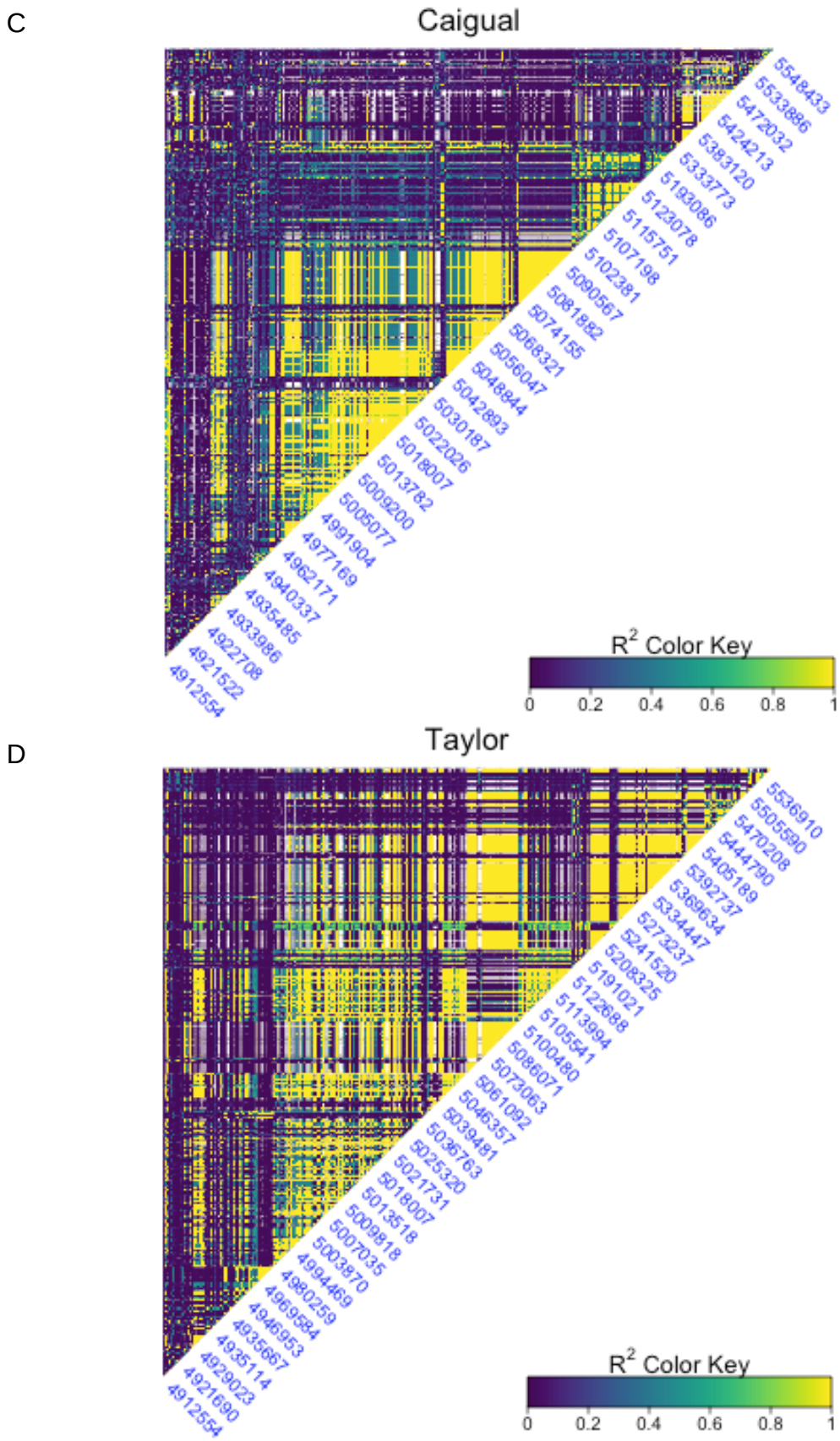


Figure 5.12 Linkage disequilibrium on chromosome 15 from 4,900,000-5,550,000 bp. A) ILL, B) IUL, C) IC, D) IT.

5.5 Discussion

In the previous chapter we showed that four rapidly evolving, experimental populations of guppies only evolved minor differences in neutral diversity compared to the HP source, and that IT showed evidence for a bottleneck after the introduction whereas the other three populations showed no evidence of bottlenecks. Here, we explored signatures of selection in these populations. We uncovered strong signals of selection in all four populations. Using a combination of haplotype genome scans and a newly developed multivariate approach, we found that some regions showed divergent selection, while others showed parallel change. Our multivariate approach was able to detect more subtle parallel changes in allele frequency compared to the genome scans. Combining the evidence, we found a region under strong selection in three of the four populations on chromosome 15, and the multivariate approach showed that all four populations are evolving in parallel in this region.

There were strong indications of recent, convergent selection on chromosome 15 for all introduced populations. This region (located around 5Mb) was previously found as a candidate for convergent evolution in long-term introduction experiments (Chapter 3) as well as in natural populations (Fraser *et al.*, 2015; Whiting *et al.*, 2020). This region also harbours genes belonging to the cadherin family, known to be involved in pigment cell migration and other cell-cell adhesion processes (Fukuzawa & Obika, 1995; Nishimura *et al*., 1999), that could be involved in the colour differences between HP and LP males. Clearly, this region is a strong candidate for adaptation to LP environments in the guppy system, and further studies such as gene knockout experiments could help identify the exact role of the cadherin genes in this process.

Rapid convergent genomic adaptation, specifically at quantitative traits, is often predicted to occur through small shifts in allele frequency (Barrett and Schluter, 2008), which are difficult to detect using most available methods. The novel multivariate allele frequency change analysis described here can identify parallel shifts in allele frequency rather than absolute changes in allele frequency at individual loci. Other existing methods, such as BayPass (Gautier, 2015), are also able to do this, however there are several features that make our approach unique. Firstly, BayPass considers all SNPs separately, whereas our multivariate approach is haplotype-based by considering all SNPs in a window together. If there is a complex haplotype structure within the window (e.g. adjacent SNPs belonging to different haplotypes), the complexity of the space is increased, which in turn reduces the level of parallelism of the region. Secondly, as a result of this, our method is able to identify different evolutionary histories within a window by interpreting the eigenvectors collectively. For example, if all population pairs show full parallelism along the primary eigenvector (suggesting selection on a common ancestral haplotype), the secondary eigenvector can provide insight in population-specific divergent haplotypes within the common ancestral haplotype. BayPass on the other hand is most effective for scenarios that are fully parallel, which is unlikely in complex datasets.

Using this newly developed method, we found 38 windows that experienced in parallel changes in allele frequency among the four populations. Most of these windows lacked a strong haplotype signal in the genome scan measures. There are several explanations that can explain this observation: first, it is possible that the variance on this axis was constrained for other reasons than selection. For example, background selection could cause correlated differentiation landscapes among the populations, which could result in a constrained axis of allele

frequency changes (Burri, 2017). Second, the strength of selection may have been relatively weak, causing time of fixation for a beneficial allele to be considerably longer. These weakly selected alleles would not have had time to generate extreme frequency differences in the introduced populations (Coop *et al.*, 2009). Finally, many of traits under selection in LP guppies are likely polygenic, therefore rapid adaptation takes place via selection on standing genetic variation at many loci simultaneously, causing subtle shifts of allele frequencies across these loci (Pritchard *et al.*, 2010). All of the factors mentioned above, would result in small parallel shifts in allele frequencies rather than the strong signals of selection left behind by (nearly) complete sweeps at single loci.

There could be other factors at play resulting in convergence at the phenotypic level but not at the genomic level. First, it is possible the evolved LP phenotypes have mainly resulted from phenotypic plasticity, which would allow them to evolve so rapidly without leaving a signature of selection at the genetic level (Lande, 2009). Indeed, Handelsman *et al*. (2013) showed that after the introduction, these populations exhibited genetic divergence as well as a plastic response for growth rate and resting metabolic rate, through comparing wild-caught and lab-reared fish. Second, the time since the introduction has likely been too short for new mutations to result in strong population divergence. This makes detecting loci with standard genome scans challenging, as many of them require substantial population differentiation for accurate detection of candidate loci under selection (Berner, 2019). Finally, as mentioned previously, it is likely the traits under selection have a complex polygenic basis. This would mean that the similar phenotypic outcome could be the result of selection on different genes in the same pathway in the different populations (Hoekstra and Nachman, 2003), resulting in signals of selection in different regions of the genome.

The experimental populations were subjected to different ecological treatments that affected their phenotypic evolution. When Reznick *et al*. (2019) investigated life history evolution in the four introduced populations, they found that populations with low initial densities (ILL and IUL) took a year longer to reach peak population densities compared to the high initial density populations (IC and IT). In Chapter 4, we showed that this longer period of population growth is likely responsible for the increased diversity in ILL and IUL compared to IC and IT. In the current chapter, we show that it also led to ILL and IUL having fewer SNPs with unchanged allele frequencies, likely because there was more time for SNPs to randomly drift away from GHP allele frequencies. We further find that IC and IT have more fixed SNPs that were the minor allele in GHP, as well as having more overlapping outlier windows in the genome scans, suggesting selection had a stronger effect in the high-density populations compared to the low-density populations.

Within each year, the canopy of one of the populations was thinned (IUL and IT). Kemp *et al*. (2018) showed that this resulted in the evolution of more blue/green iridescence in male guppies, and Reznick *et al*. (2019) found that the populations in thinned canopy sites experienced stronger seasonal fluctuations in population densities. In the previous chapter we found no differences at the genome-wide level for these different ecological treatments, and in this chapter again we saw no effect of canopy treatment on the selected alleles. This could suggest that the rapid phenotypic responses of the populations to the canopy changes have mostly been through phenotypic plasticity, or the genetic response to the ecological changes is slower than the phenotypic response (Price *et al.*, 2003). By tracking these populations over time, we could assess whether rapid

phenotypic adaptation to these ecological differences is followed by slower genetic assimilation.

Previous studies found that IT had diverging patterns of growth and evolution compared to the other populations. For example, Torres Dowdall *et al.* (2012) and Fitzpatrick *et al.* (2014) observed a higher mortality in IT during the first year which they attributed to indicators of disease found in this population but not in the other three. Similarly, Reznick *et al.* (2019) observed a major population crash in IT during the first year. In Chapter 4, we found evidence of a bottleneck and reduced diversity in IT. Based on these results, it could be expected that IT might have experienced a stronger selection pressure compared to the other populations or selection on a different set of variants (i.e. strong immune responses to overcome disease), and as a result exhibit different signals of selection than ILL, IUL and IC. However, we find no strong evidence that this is the case. Average allele frequency changes were similar across the four populations, and in our overlapping outlier analysis, comparisons including IT consistently had high numbers of overlapping outlier windows (figure 5.4) suggesting it experiences similar selection to the other populations. Rather comparisons with IUL had the lowest overlapping windows in the outlier analysis, which raises interesting hypotheses about the role of drift in this population that would require further investigation.

In conclusion, we found strong signals of selection in all four populations, and a region on chromosome 15 contained overlapping windows with strong signals of selection in three of the four populations. Further analysis of this region showed that all four populations experienced parallel allele frequency changes. Using our adapted multivariate analysis of allele frequencies, we were able to find evidence of subtle parallel changes in allele frequency that were not observed with the

standard genome scans. This suggests this method has the potential to detect convergent evolution in rapidly evolving populations, as well as identifying signatures of polygenic selection with small changes at many loci simultaneously. It would be interesting to keep investigating these populations in the future to further our understanding of the process of rapid adaptation in a model system. However, the methods developed and used here could also be applied beyond the guppy system, to investigate genomic adaptation in other rapidly changing environments.

Chapter 6 - GENERAL DISCUSSION

It is becoming increasingly clear that natural ecosystems are experiencing rapid changes as a result of human activities (Rockström *et al.*, 2009), therefore it is important we understand how species and populations cope with these rapid changes. Many studies have now shown that populations can phenotypically adapt to sudden environmental shifts on an ecological timescale of just a few generations. However, the genetic basis of this rapid adaptation still remains largely unknown. In this thesis, I aimed to examine the genomic basis of rapid adaptation to a novel environment. I used whole genome sequencing data from experimentally introduced populations of guppies to infer demographic histories and identify signatures of (convergent) evolution across the genome. Specifically, my objectives were to:

a) Investigate the impact of the introductions on genetic diversity of the study populations (Chapter 2 & 4).
b) Study how initial genetic variation affects selection and the occurrence of molecular convergence (Chapter 3).
c) Analyse the signatures of selection in the early stages of adaptation (Chapter 5).

In Chapter 2, I showed how initial genetic variation varies after translocation using two well-established experimental populations. Turure (TULP), the oldest population, had signals of high initial genetic variation, and experienced extensive population growth after the introduction. El Cedro (ECLP) on the other hand, was introduced more recently from an already bottlenecked population, and the introduction led to a further reduction of genetic diversity and a smaller effective population size.

In Chapter 3, I found strong signals of within-population selection in both established experimental populations, suggesting the limited amount of initial variation in ECLP did not affect the selective potential of this population. I also found that, even though both introduced populations showed signatures of convergent evolution with a natural LP population, there was little evidence for convergence between the experimental populations. This suggests that the convergent phenotypes of these populations were obtained through different molecular approaches.

In Chapter 4, I showed that four recently introduced populations had only minor genetic differences from their HP source, and that the shallow history of these populations likely caused problems when inferring their demographic histories based on SFS. Using runs of homozygosity, however, I uncovered no evidence of bottlenecks in three of the four populations (Lower Lalaja (ILL), Upper Lalaja (IUL) and Caigual (IC)), whereas Taylor (IT) did show signs of a bottleneck.

In the last data chapter (Chapter 5), I uncovered strong signals of selection in all four populations. Specifically, a 2Mb region on chromosome 15 showed strong convergent selection in three of the four populations. Using a newly developed multivariate approach, I revealed subtle parallel changes in allele frequency across all populations, including the candidate region on chromosome 15, that remained undetected by the standard genome scans.

In this final chapter, I will review these findings in relation to my objectives and the questions I outlined in chapter 1, discuss the strengths and limitations of the research, and highlight opportunities for further research.

6.1 The importance of demography

It is predicted that introduced species would experience a founding bottleneck after being translocated to a new environment, resulting in reduced genetic diversity, however there is increasing evidence from observations in nature that not all introduced populations exhibit evidence of genetic bottlenecks (Dlugosch and Parker, 2008). Using whole genome sequencing data, I inferred the demographic histories of experimentally introduced populations of guppies and showed that indeed translocations do not necessarily result in detectable bottlenecks and reduced genetic diversity. The two established introduction populations represented two very different demographic scenarios, and revealed that a limited amount of initial variation at the start of the experiment was followed by a further reduction of genetic variation in ECLP but not in TULP. In the recently introduced populations, I found no evidence for founding bottlenecks in three of the four populations. Combined with previous research on census size of these populations (Torres-Dowdall, Handelsman, Ruell, *et al.*, 2012; Fitzpatrick *et al.*, 2014), the bottleneck in the fourth population (IT) likely occurred sometime during the first year, as a result of increased mortality caused by disease and extreme seasonal fluctuations of population size.

Other studies examining invasive and human-introduced population genetics also found that increased initial genetic diversity in introduced populations allowed them to avoid founding bottleneck effects (Wares *et al.*, 2005; Roman and Darling, 2007). In many of these studies however, the increased genetic diversity was attributed to multiple introductions providing additional genetic diversity (Roman and Darling, 2007). In the populations investigated in this thesis, this was not the case, as we know only one introduction event took place and migration upstream is extremely limited in the Northern Range Mountains (Willing *et al.*, 2010; Whiting *et al.*, 2020). Instead, the

large number of individuals introduced in each experiment will likely have limited the effects of random genetic drift after the introduction event, (Charlesworth, 2009). Furthermore, in the case of the recently introduced populations, the experimental approach was designed to increase the effective population size at the time of the introduction. Guppies were mated in groups of five males and five females, and subsequently released into the experimental sites with fish from other breeding groups. Female guppies still carried sperm from the males of their breeding group when they were introduced with a new group of males in each respective river, increasing the effective population size. Finally, several aspects of the life histories of guppies, such as the fast generation time and multiple mating of females will help to maintain high genetic variation (Becher and Magurran, 2004).

While the evidence for the effect of limited initial genetic variation on the likelihood of molecular convergence in this thesis comes from only one population (ECLP), the results provide insight into a potential mechanism determining the genetic response of an introduced population and support the above-mentioned example. It also shows that the success of an introduction is not necessarily limited by low genetic diversity and high levels of inbreeding. Previous studies showed that introduced ECLP fish evolved LP life histories (Reznick and Bryga, 1987), and more elaborate colouration compared to HP fish (Kemp *et al.*, 2009), and my results found strong signals of selection across the genome, suggesting that limited variation still allowed for selection and rapid adaptation. A similar result was found by a study investigating genetic variation in re-introduced populations of the Common hamster (La Haye *et al.*, 2017). They showed that populations were successfully established from a low-genetic diversity breeding line, although the populations experienced a further decline

in genetic diversity after the introduction, whereas populations established from multiple breeding lines with higher genetic diversities retained genetic diversity.

6.2 Molecular convergence

One of the major questions in the field of evolutionary biology remains whether phenotypic convergence is underpinned by molecular convergent evolution. The guppy system in Trinidad provides an excellent model to study this question because of the strong convergent phenotypes observed in LP populations across the Northern Range Mountains. Previously, Fraser *et al.* (2015) investigated molecular convergent evolution using a RAD-seq approach in two experimental populations (one of which was ECLP) and three natural LP populations. They found the two introduced populations shared more outlier regions than would be expected by chance, but found very few shared regions across all populations. Similarly, Whiting *et al.* (2020) used whole genome sequencing data of five natural HP-LP pairs and found limited evidence for molecular convergence among all population pairs but a strong convergent candidate within the Caroni drainage.

In Chapter 3, I showed that the bottlenecks in the El Cedro river likely removed SGV, resulting in little shared SGV with TULP, and therefore a different set of genetic variants available for selection to act upon. As a consequence, despite both long-term experimental populations (ECLP and TULP) showing molecular convergent evolution with a natural LP population, there was little evidence for molecular convergence between the experimental populations. A region worth mentioning is the convergent outlier region between GLP/GHP and TULP/GHP on chromosome 20/scaffold 94 (Chapter 3). Whiting *et al.* (2020) describe a large, (nearly) fixed haplotype in three LP Caroni populations making this region an excellent candidate for adaptation. Whiting

et al. then go on to speculate that this region was lost in the Oropuche drainage resulting in limited convergent evolution across the whole guppy distribution. Investigating the haplotype structure in TULP (based in the Oropuche drainage but introduced from the Caroni drainage) revealed that indeed, it lacked the haplotype structure observed in the populations from the Caroni (GHP, GLP, ECHP, and ECLP). This creates an exciting framework to investigate the importance of introgression in adaptation by focussing on the relationship between TULP (Caroni origin but introgressing with Oropuche) and a Turure HP population (Oropuche origin but introgressing with Caroni).

The distinct evolutionary trajectories I observed in the established populations show how convergence at the phenotypic level does not necessarily reflect convergence at the molecular level. It is possible however that convergence could be observed at higher levels of biological organisation, for example if different genes under selection are part of the same genetic pathway. Jacobs *et al.* (2020) found a similar pattern in Arctic charr, where they observed extensive convergence of phenotypes and gene expression patterns, but few shared outlier regions among the ecotypes.

In the recently introduced populations, I uncovered 38 windows experiencing parallel allele frequency changes across all four populations, using a newly developed multivariate analysis of allele frequencies (Chapter 5). Additionally, when I overlapped these windows with the outlier windows identified by the haplotype genome scans, I identified a region on chromosome 15 (located between 4 - 5 Mb) with strong signals of selection in three of the four populations (ILL, IC and IT). The region contains genes from the cadherin family that are known to be involved in pigment cell migration (Fukuzawa & Obika, 1995; Nishimura et al., 1999), however it is not yet known what

the exact function of the region is, and future studies could work to identify the functional aspect of these genes in the development and evolution of LP guppies.

To examine any commonalities in outlier windows among all populations I overlapped the outlier lists of TULP and ECLP (supplementary tables S3.1 and S3.2) with the identified outlier windows on eigenvector 1 (supplementary table S5.9). For TULP and the new introduction populations, there were two regions on chromosome 15 that overlapped (table 6.1), discussed above. This region was also identified as an outlier region in natural populations (Fraser *et al.*, 2015; Whiting *et al.*, 2020), and thus provides strong evidence for molecular convergence across many populations. For ECLP and the younger populations I found one overlapping window on chromosome 10 that contained two genes (table 6.1). One of these genes (Casein kinase I isoform alpha, csnk1a1) is involved in the phosphorylation of cadherin-associated protein β (ctnnb1), which in turn functions as a key component of the cadherin complex by forming adherens junctions with cadherin 1 (cdh1) (Nelson and Nusse, 2004; Wang *et al.*, 2009).

Table 6.1 Overlapping outlier regions among TULP and the new introduction populations and among ECLP and the new introduction populations

		Eigenvector 1 outlier		
Population	**Chromosome**	**BP1**	**BP2**	**Genes**
TULP	chr15	4938877	4985391	loc10347655, cdh1
		5021202	5060367	B-cadherin, cdh1
ECLP	chr10	1044123	1141168	csnk1a1, fbxo38

It could be expected that the younger introduction populations would exhibit higher levels of molecular convergence as they all originated from the same source population and thus share the same SGV (Conte *et al.*, 2012). In chapter

5, I showed that, although all populations showed strong signals of selection, overlap of outlier windows was limited among the experimental populations. By investigating allele frequency changes across the genome however, I found that the introduced populations showed a high incidence of small parallel changes in allele frequencies across the genome. A similar combination of convergent polygenic shifts across all populations, but limited convergence among strong signals of selection was observed in rapidly evolving populations of Atlantic silversides (Therkildsen *et al.*, 2019).

Other studies investigating rapid genomic convergence in wild populations found higher levels of molecular convergence than I did here (Bassham *et al.*, 2018; Marques *et al.*, 2018; van Boheemen and Hodgins, 2020). The difference with these studies could lie in the strength of the population bottlenecks in these populations. For example, Van Boheemen and Hodgins (2017; 2020) identified a bottleneck in an introduced population of *Ambrosia artemisiifolia*, but found that neutral genetic diversity had remained relatively high. This suggests the bottleneck was not very strong and as a result, the population retained most of its shared SGV with the source and a second introduced population that had not experienced a bottleneck. This explains the lack of convergence between the bottlenecked populations (ECLP and IT) and the other populations, but not for TULP, ILL, IUL and IC. Whiting *et al.* (2020) attribute the lack of molecular convergence in natural populations to the strict structuring of genetic variation between rivers, resulting in limited shared genetic variation across the rivers.

6.3 The impact of time on measuring genomic adaptation

Throughout this thesis, I focussed on sets of experimental populations that were established at different times in the past. All three experiments used source

populations from the same drainage, either from the Guanapo river (TULP, ILL, IUL, IC and IT) or a close tributary of the Guanapo (El Cedro). The main difference between the populations was the time since divergence from their HP source. The oldest population (TULP) was introduced in 1957 (114 generations), followed by ECLP in 1980 (64 generations) and finally ILL, IUL, IC and IT in 2008/9 (8-10 generations). Here, I will bring together the results from these experiments to analyse the effect of time since divergence on our ability to measure genomic adaptation.

I found signals of genomic adaptation and selection in all six populations, but there were differences in the magnitude of these signals. Population differentiation (measured as F_{ST}) was highest in ECLP/ECHP (median F_{ST}=0.061), followed by TULP/GHP (median F_{ST}=0.056) and the newly introduced populations had much lower values (median F_{ST} ranged from 0.013 to 0.022). The number of outlying XP-EHH windows followed a similar pattern, with most windows identified in ECLP/ECHP (N=288), followed by TULP/GHP (N=118), whereas the recent introductions ranged from 17 to 57 outlier windows. This suggests there were more (nearly) complete sweeps at single loci in the older pops (Coop *et al.*, 2009). This pattern is slightly surprising, because based on the ages of the populations, it could be expected that ECLP values of each measure should fall in between the TULP values and newly introduced values. However, the bottlenecked history of ECHP has caused this HP/LP pair to become an outlier that has its own patterns of genomic adaptation and selection, making it difficult to compare the results of the ECLP experiment to the other populations.

Time since divergence can also play a role in the amount of molecular convergence we observe between the oldest population and the newly introduced

populations. For example, different traits could be selected upon in different stages of adaptation. In a long-term experiment investigating evolutionary dynamics in *Escherichia coli*, Good *et al.* (2017) indeed found that the targets of selection shift over time. They explain that changing ecological interactions and novel mutations generate genetic opportunities for selection to act upon that were initially not available. For example, selection could change over time as a result of increasing/decreasing population density. Previous studies in the guppy system have shown that population density plays an important role in the process of adaptation (Reznick, Bassar, *et al.*, 2019). It is possible that TULP is experiencing selection in a region that is not yet beneficial in the younger populations. The recently introduced populations are still accessible for data collection, making it possible to track shifting patterns of signatures of selection over time and investigate how these shifts affect the occurrence of molecular convergence.

In this thesis, I discovered that time since divergence affected my ability to infer demographic histories from populations in different stages of adaptation. I found that the newly introduced populations were not substantially genetically differentiated from their HP source, and therefore their demographic histories reflected their ancestor's (GHP) history rather than their own. If any new mutations have arisen during the 8 to 10 generations since the introduction, they would only be present at low frequencies, and in order to detect these rare variants we would need much larger sample sizes (N>100) (Beichman *et al.*, 2018). As a result, I was not able to identify a single best model for three of the four populations, and in the fourth (ILL) the simulated parameter distributions were unrealistic and did not match the observed data. In contrast, for the two

older experiments, model selection using the same method resulting in one likely model, and parameter estimates were comparable with observed data.

Similarly, Elleouet and Aitken (2018) reported that ABC methods were less accurate when inferring parameters of recent demographic events, and Benazzo *et al.* (2015) struggled to obtain parameter estimates for populations with low levels of genetic differentiation using both (an earlier version of) FASTSIMCOAL2 and ABC methods. In an introduction experiment with threespine stickleback however, Marques *et al.* (2018) quantified genomic change ~13 generations after the transplant and successfully inferred demographic histories using FASTSIMCOAL2. However, the F_{ST} value they report is much higher than those found in the recently introduced populations in this thesis, indicating the introduced population was more differentiated from its source than our newly introduced populations. Other studies applying this method also report high values of F_{ST} and generally consider much older populations (Bagley *et al.*, 2017; Raposo do Amaral *et al.*, 2018).

The range of experimental time points in this thesis allowed me to investigate appropriateness of available methods for detecting selection and inferring demographic histories in populations of different ages. I showed that demographic inference using the SFS is possible for populations established over 60 generations ago, but was problematic for populations younger than 15 generations. Whether demographic inference from the SFS is appropriate also depends on the level of differentiation between the source and the introduced populations. One promising new method infers demographic histories based on the observed spectrum of LD of pairs of loci in a contemporary population sample, allowing for accurate estimations of changes in population sizes, even in the first 20 generations (Santiago *et al.*, 2020). Other methods used in this thesis, such

as runs of homozygosity and haplotype-based genome scans were found to be appropriate for populations of all ages. Finally, the results in Chapter 5 revealed that rapid genomic adaptation likely occurred through small shifts in allele frequency at multiple loci simultaneously, a signal of polygenic selection. It would thus be advisable to explore methods considering frequency shifts of alleles simultaneously at all loci involved (instead of individual SNPs) (Jain and Stephan, 2017).

To conclude, I have shown that demographic histories and the age of a population play a major role in the likelihood of molecular convergence. Because both these factors vary extensively between populations, the underlying genetic basis of adaptation is much more variable than we can currently comprehend from the limited number of studies investigating molecular convergence in a variety of species. The work presented here showed that in the case of rapid adaptation, genomic convergence is likely the result of small, parallel allele frequency shifts at multiple loci simultaneously. These results highlight that, in order to further our understanding of the genomic basis of rapid adaptation new methods are necessary that can detect these subtle parallel shifts in WGS data. Our multivariate analysis of allele frequency changes approach provides a promising avenue for detecting polygenic signals of selection.

6.4 Confounding environmental variables

Using populations in the wild in any research requires an acknowledgement of the unmeasured and unknown. Natural ecosystems are complex and even though studying wild populations is essential for producing real-world data, this complexity also creates noise and uncertainty that complicate the interpretation of results (Garland Jr. and Rose, 2009). Within a local site, many ecological

factors are unavoidably correlated with each other (e.g. predation pressure and population density in guppy populations (Reznick, Bassar, *et al.*, 2019)). These correlations make it difficult to tease apart the relative impacts of ecological factors on the population under investigation. If a correlation between predation and a certain environmental factor (variable X) exists, it is possible the observed signals of selection are actually associated with variable X rather than predation pressure. Furthermore, the covariance of ecological factors can vary between sites. For example, in one river predation always correlates with variable X, but in a second river no such correlation exists. This creates a divergent selection regime within multivariate space.

Additionally, ecological factors undoubtedly show quantitative environmental differences among habitats that are considered qualitatively similar (Stuart *et al.*, 2017). For example, streams in the Trinidadian guppy system are generally regarded as discrete classes of high- and low predation, whereas actually predation pressure varies within and among rivers and drainages (Deacon *et al.*, 2018). Therefore, what researchers consider replicate populations might in fact not be as strictly replicated as populations in a fully controlled (laboratory-based) experiment, and this limits our ability to identify incidences of molecular convergent evolution. Thompson *et al*. (2019) reported that even small deviations in environment can drastically reduce the fraction of alleles that are beneficial in both populations, and therefore reduce the amount of convergence among populations.

Although these correlations and slightly varying habitats can confound the interpretation of results, they also make studying *in situ* experiments interesting and provide data that is more representative of real life than laboratory experiments. In a laboratory setting, selection is often imposed on a single focal

factor by creating a simple environment and eliminating sub-optimal conditions (Kawecki *et al.*, 2012). As a result, the patterns of selection observed in these experiments are likely too idiosyncratic of the laboratory environment and not very informative for understanding genetic adaptation in natural populations. So, even though laboratory-based experimental evolution has contributed significantly to our understanding of the process of genetic adaptation (Kawecki *et al.*, 2012; Blount *et al.*, 2018), it is essential we expand experimental evolution to include more investigations of natural populations. By doing so we can better understand how populations respond to environmental change, which could have important implications for conservation and ecosystem functioning (Collins, 2011).

In Chapter 4 and 5, I highlight the unique replicated nature of the recently introduced populations as an advantage for investigating the early stage of genomic adaptation. However, with the various ecological treatments the populations have been given (table 6.2) it could be argued that, technically we had no true replicate populations, and as a result there is less molecular convergence among the populations. De Lisle and Bolnick (2020) showed that by analysing the spectral decomposition of covariance matrices among lineages, it is possible to reveal shared dimensions of deterministic evolutionary change, even in not fully parallel environments. Using an adapted version of their method to analyse parallel shifts in allele frequency changes, I found that indeed, these populations show a considerable amount of molecular convergence. This suggests that despite the ecological differences, these populations are still overall evolving convergently.

Table 6.2 Ecological treatments per population in newly introduced populations

	Low density	**High density**
Open canopy	IUL	IT
Closed canopy	ILL	IC

In an ideal world, with unlimited resources, it would be interesting to perform an exhaustive study that encompasses interactions among multiple ecological aspects, replicated multiple times and investigate how those interactions influence genomic adaptation. By having multiple replicates for each ecological scenario, it would be possible to investigate the effect of the various ecological aspects on the genomic adaptation of guppies.

6.5 Genomic context and the search for signals of selection

Increasingly more attention has been paid to the genomic context in the search for signals of selection. Heterogeneous distribution of intrinsic genomic properties, such as recombination and mutation rate, can affect the way natural selection impacts levels of variation across the genome (Stankowski *et al.*, 2019). Regions of reduced recombination can maintain beneficial combinations of alleles (Dobzhansky, 1937), which can lead to increased occurrences of molecular convergence. For example, regions of low recombination contributed to the convergence of social chromosomes across ant species (Purcell *et al.*, 2014). Low-recombination regions have also been implicated in repeated adaptation in stickleback (Samuk *et al.*, 2017) and cichlids (Meier *et al.*, 2018). Burri *et al.* (2015) also warn that convergent patterns of high differentiation in closely related taxa could be the result of linked selection in low-recombination

regions if the recombination landscape among the taxa is conserved. This has shown to especially affect taxa with highly stable genome landscapes, such as birds (Burri *et al.*, 2015; Singhal *et al.*, 2015), sunflowers (Renaut *et al.*, 2013) and *Heliconius* butterflies (Martin *et al.*, 2019). The recombination landscape of the guppy has not been extensively studied. Charlesworth *et al.* (2020) recently investigated intronic GC content to indirectly infer the recombination patterns on guppy chromosomes and found patterns of GC peaks were similar across guppy populations, and two closely related species (*Xiphophorus maculatus* and *Poecilia picta*).

Just as recombination rate can cause signals of molecular convergence, so can mutation bias. Mutations are a critical source of genetic variation, therefore how often they occur, where they occur, and the types of mutations (mutation spectrum) will affect the pool of available genetic variation (Sane *et al.*, 2020). A strong mutational bias can influence the outcome of the evolutionary process and thus determine the genetic basis of adaptation and drive convergent evolution (Stoltzfus and McCandlish, 2017; Storz *et al.*, 2019). Mutation bias is less likely to affect the youngest introduced populations I investigated here, as not enough time has passed for new mutations to occur. In the older introductions however, it is possible localised mutation rates have caused correlated patterns of differentiation. Therefore, it is possible the genomic architecture of the guppy genome could result in correlated patterns of differentiation. It would therefore be beneficial to further investigate the patterns of these intrinsic properties in the guppy genome and incorporate them in our overlapping outlier approach as this would allow us to distinguish signatures of local adaptation from those resulting from underlying genome features.

Throughout the thesis I focussed solely on SNPs, however, structural variants (SV) could also be involved rapid adaptation. SVs exist in the form of copy number variants (CNV), inversions, and insertions and deletions. It has been suggested that SV's are the largest source of standing variation across the genome, and can explain much more phenotypic variation than SNPs (Zhou *et al.*, 2019). Several studies have already successfully identified SV's underlying phenotypic variation in other species (Le Moan *et al.*, 2020; Zhou and Gaut, 2020). Identifying SVs is problematic however, as SVs can be very large and cover most of a sequencing read or even be larger than the read length, making mapping difficult (Sedlazeck *et al.*, 2018). Furthermore, multiple SVs can overlap or be nested, creating complex mapping patterns that may impede mapping altogether. SVs should be correlated with SNPs, however, at least at their boundaries, and therefore we would have been able to detect SNP variation correlated with SVs using our window-based analyses. For example, in TULP we found a large region on chromosome 20/scaffold 94 (about 3.5Mb) that showed strong patterns of selection across the entire section. This pattern is suggestive of a chromosomal inversion, although Whiting *et al.* (2020) did not find evidence for SVs in the region using short read technology. Recent methodological and technological developments, such as long-read sequencing, will help to increase the detection of SVs (Mahmoud *et al.*, 2019) and might therefore be a fruitful area of future investigations of SVs in the guppy genome.

6.6 Other avenues for future research

Throughout the last five years, the field of evolutionary genetics continued to develop rapidly. The tempo at which new methods and ideas appear makes it challenging to project into the future, however some trends are likely to provide

avenues for further research. For example, the availability of low-cost whole genome sequencing data from non-model organisms will continue to increase, as well as the availability of low-coverage computational advances (e.g. Angsd (Korneliussen *et al.*, 2014)). These improvements will make it possible to obtain hundreds to thousands of whole genomes per population in the near future. With such large sampling sizes, methods based on the site frequency spectrum should become more powerful, as rare variants can be more reliably detected. This would allow for more accurate inference of very recent population histories such as those from recently introduced or invasive species.

Another emerging trend is the use of temporal data to investigate the genomic architecture of adaptation. The recently introduced replicate populations provide a unique opportunity to obtain and use temporal genomic data. By tracking these populations over time, we could greatly improve our understanding of the evolutionary dynamics of rapid genomic adaptation in natural populations. Buffalo and Coop (2019, 2020) recently described a temporal approach that analyses temporal covariances between allele frequency changes from one timepoint to the next, allowing for the detection of the genome-wide impact of subtle allele frequency shifts associated with selection on SGV and polygenic traits. Using temporal genomic data, it would also be possible to investigate fluctuations in direction and strength of selection over time, such as seasonal effects (Bergland *et al.*, 2014; Gompert, 2020). For example, all the introduced populations showed strong seasonal fluctuations of population density (especially in the thinned-canopy populations) (Reznick, Bassar, *et al.*, 2019) and it would be interesting to investigate whether these fluctuations impose temporally variable selection on these populations. These experiments are still accessible and we can study them as they continue to evolve.

6.7 Concluding remarks

In this thesis, I examined patterns of rapid genomic adaptation in several experimentally introduced populations of guppies, and shown how these patterns were affected by demographic histories. My analysis of rapid genomic adaptation in the earliest stages of adaptation is one of the first in the field to do so for replicated experimental populations in the wild. This study highlighted that rapid genomic adaptation occurs through small, parallel shifts of allele frequencies in multiple loci, and that the limited time for genetic variation to arise can cause a problem for inferring demographic histories. I have also shown that historical bottlenecks reduced the amount of shared SGV between populations, resulting in limited evidence for molecular convergence despite convergent phenotypes. Finally, a region on chromosome 15 (located at 5Mb) was identified as a strong candidate for molecular convergence across many populations in the Trinidadian guppy system.

Altogether, my results show that evolution of convergent phenotypes is repeatable at the genomic level, but that extent of molecular convergence is associated with factors such as shared standing genetic variation and time since divergence. The findings in this thesis will most likely apply to rapidly evolving populations and species outside the Trinidadian guppy system, and therefore hopefully provide a useful guide and source of information for future researchers with an interest in rapid genomic adaptation.

REFERENCES

Alkins-Koo, M. (2000) Reproductive timing of fishes in a tropical intermittent stream, *Environmental Biology of Fishes*, 57(1), pp. 49–66. doi: 10.1023/A:1007566609881.

Anderson, J. T., Panetta, A. M. and Mitchell-Olds, T. (2012) Evolutionary and ecological responses to anthropogenic climate change, *Plant Physiology*, 160(4), pp. 1728–1740. doi: 10.1104/pp.112.206219.

Andreev, D., Kreitman, M., Phillips, T. W., Beeman, R. W. and ffrench-Constant, R. H. (1999) Multiple Origins of Cyclodiene Insecticide Resistance in Tribolium castaneum (Coleoptera: Tenebrionidae), *Journal of Molecular Evolution*, 48(5), pp. 615–624. doi: 10.1007/PL00006504.

Andrews, S. (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available at: www.bioinformatics.babraham.ac.uk/projects/fastqc.

Arendt, J. D. and Reznick, D. N. (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation?, *Trends in Ecology and Evolution*, 23(1), pp. 26–32. doi: 10.1016/j.tree.2007.09.011.

Arendt, J. D., Reznick, D. N. and López-Sepulcre, A. (2014) Replicated origin of female-biased adult sex ratio in introduced populations of the Trinidadian guppy (Poecilia reticulata), *Evolution*, 68(8), pp. 2343–2356. doi: 10.1111/evo.12445.

van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Levy-moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. and Depristo, M. A. (2014) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline, *Current Protocols in Bioinformatics*, 11(1110). doi: 10.1002/0471250953.bi1110s43.From.

Bagley, R. K., Sousa, V. C., Niemiller, M. L. and Linnen, C. R. (2017) History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (Neodiprion lecontei), *Molecular Ecology*, 26(4), pp. 1022–1044. doi: 10.1111/mec.13972.

Baldwin, J. M. (1896) A new factor in evolution, *The American naturalist*, 30, pp. 441–451. doi: 2453130.

Barrett, R. D. H. and Schluter, D. (2008) Adaptation from standing genetic variation, *Trends in Ecology and Evolution*, 23(1), pp. 38–44. doi: 10.1016/j.tree.2007.09.008.

Barson, N. J., Cable, J. and Van Oosterhout, C. (2009) Population genetic analysis of microsatellite variation of guppies (Poecilia reticulata) in Trinidad and Tobago: Evidence for a dynamic source-sink metapopulation structure, founder events and population bottlenecks, *Journal of Evolutionary Biology*, 22(3), pp. 485–497. doi: 10.1111/j.1420-9101.2008.01675.x.

Barton, N. H. and Keightley, P. D. (2002) Understanding quantitative genetic variation, *Nature Reviews Genetics*, 3(1), pp. 11–21. doi: 10.1038/nrg700.

Bassham, S. L., Catchen, J., Lescak, E., von Hippel, F. A. and Cresko, W. A. (2018) Repeated selection of alternatively adapted haplotypes creates sweeping genomic remodeling in stickleback, *Genetics*, 209(July), p. genetics.300610.2017. doi: 10.1534/genetics.117.300610.

Beatty, J. (1984) Chance and natural selection, *Philosophy of Science*, 51, pp. 183–211.
Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate Bayesian Computation in population genetics, *Genetics*, 162(4), pp. 2025–2035. doi: 10.1111/j.1937-2817.2010.tb01236.x.
Becher, S. A. and Magurran, A. E. (2000) Gene flow in Trinidadian guppies, *Journal of Fish Biology*, 56(2), pp. 241–249. doi: 10.1111/j.1095-8649.2000.tb02103.x.
Becher, S. A. and Magurran, A. E. (2004) Multiple mating and reproductive skew in Trinidadian guppies, *Proceedings of the Royal Society B: Biological Sciences*, 271(1543), pp. 1009–1014. doi: 10.1098/rspb.2004.2701.
Beichman, A. C., Huerta-Sanchez, E. and Lohmueller, K. E. (2018) Using genomic data to infer historic population dynamics of nonmodel organisms, *Annual Review of Ecology, Evolution, and Systematics*, 49(1), pp. 433–456. doi: 10.1146/annurev-ecolsys-110617-062431.
Benazzo, A., Ghirotto, S., Vilaça, S. T. and Hoban, S. (2015) Using ABC and microsatellite data to detect multiple introductions of invasive species from a single source, *Heredity*, 115(3), pp. 262–272. doi: 10.1038/hdy.2015.38.
Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S. and Petrov, D. A. (2014) Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila, *PLoS Genetics*, 10(11). doi: 10.1371/journal.pgen.1004775.
Berner, D. (2019) Allele Frequency Difference AFD – an intuitive alternative to FST for quantifying Genetic population differentiation, *Genes*, 10(308).
Bertolini, F., Cardoso, T. F., Marras, G., Nicolazzi, E. L., Rothschild, M. F. and Amills, M. (2018) Genome-wide patterns of homozygosity provide clues about the population history and adaptation of goats, *Genetics Selection Evolution*, 50(1), pp. 1–12. doi: 10.1186/s12711-018-0424-8.
Blischak, P. D., Barker, M. S. and Gutenkunst, R. N. (2020) Inferring the demographic history of inbred species from genome-wide SNP frequency data, *Molecular Biology and Evolution*, 37(7), pp. 2124–2136. doi: 10.1093/molbev/msaa042/5739971.
Blount, Z. D., Lenski, R. E. and Losos, J. B. (2018) Contingency and determinism in evolution: Replaying life's tape, *Science*, 362(6415), p. eaam5979. doi: 10.1126/science.aam5979.
van Boheemen, L. A. and Hodgins, K. A. (2020) Rapid repeatable phenotypic and genomic adaptation following multiple introductions, *Molecular Ecology*, pp. 1–16. doi: 10.1111/mec.15429.
van Boheemen, L. A., Lombaert, E., Nurkowski, K. A., Gauffre, B., Rieseberg, L. H. and Hodgins, K. A. (2017) Multiple introductions, admixture and bridgehead invasion characterize the introduction history of Ambrosia artemisiifolia in Europe and Australia, *Molecular Ecology*, 26(20), pp. 5421–5434. doi: 10.1111/mec.14293.
Bradshaw, A. D. (2006) Unravelling phenotypic plasticity – why should we bother?, *New Phytologist*, 170(4), pp. 6–17. doi: 10.1111/j.1469-8137.2006.01758.x.
Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. and Stephan, W. (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms, *Genetics*, 140(2), pp. 783–796. Available at: http://www.genetics.org/content/genetics/140/2/783.full.pdf (Accessed: 12 June 2017).
Broad Institute (2018) Picard Toolkit. Available at:

http://broadinstitute.github.io/picard.
Browning, S. R. and Browning, B. L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, *American Journal of Human Genetics*, 81(5), pp. 1084–1097. doi: 10.1086/521987.
Browning, S. R. and Browning, B. L. (2015) Accurate non-parametric estimation of recent effective population size from segments of identity by descent, *American Journal of Human Genetics*, 97, pp. 404–418. doi: 10.1016/j.ajhg.2015.07.012.
Brüniche-Olsen, A., Kellner, K. F., Anderson, C. J. and DeWoody, J. A. (2018) Runs of homozygosity have utility in mammalian conservation and evolutionary studies, *Conservation Genetics*, 19(6), pp. 1295–1307. doi: 10.1007/s10592-018-1099-y.
Buffalo, V. and Coop, G. (2019) The linked selection signature of rapid adaptation in temporal genomic data, *Genetics*, 213(3), pp. 1007–1045. doi: 10.1534/genetics.119.302581.
Buffalo, V. and Coop, G. (2020) Estimating the genome-wide contribution of selection to temporal allele frequency change, *Proceedings of the National Academy of Sciences*, 117(34), pp. 1–9. doi: 10.1101/798595.
Burri, R. (2017) Interpreting differentiation landscapes in the light of long-term linked selection, *Evolution Letters*, 1(3), pp. 118–131. doi: 10.1002/evl3.14.
Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., Suh, A., Dutoit, L., Bureš, S., Garamszegi, L. Z., Hogner, S., Moreno, J., Qvarnström, A., Ružić, M., Sæther, S. A., Sætre, G. P., Török, J. and Ellegren, H. (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers, *Genome Research*, 25(11), pp. 1656–1665. doi: 10.1101/gr.196485.115.
Carroll, S. P. and Boyd, C. (1991) Host race radiation in the soapberry bug: natural history with the history, *Evolution*, 46(4), pp. 1052–1069.
Castric, V., Bonney, F. and Bernatchez, L. (2001) Landscape structure and hierarchical genetic diversity in the brook charr, Salvelinus fontinalis, *Evolution*, 55(5), pp. 1016–1028. doi: 10.1554/0014-3820(2001)055[1016:lsahgd]2.0.co;2.
Catullo, R. A., Llewelyn, J., Phillips, B. L. and Moritz, C. C. (2019) The potential for rapid evolution under anthropogenic climate change, *Current Biology*, 29, pp. R996–R1007. doi: 10.1016/j.cub.2019.08.028.
Ceballos, F. C., Hazelhurst, S. and Ramsay, M. (2019) Runs of homozygosity in sub-Saharan African populations provide insights into complex demographic histories, *Human Genetics*, 138(10), pp. 1123–1142. doi: 10.1007/s00439-019-02045-1.
Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. and Wilson, J. F. (2018) Runs of homozygosity: Windows into population history and trait architecture, *Nature Reviews Genetics*. Nature Publishing Group, pp. 220–234. doi: 10.1038/nrg.2017.109.
Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. and Lee, J. J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets, *GigaScience*, 4(1), p. 7. doi: 10.1186/s13742-015-0047-8.
Charlesworth, B. (2009) Effective population size and patterns of molecular evolution and variation, *Nature Reviews Genetics*, 10(3), pp. 195–205. doi: 10.1038/nrg2526.
Charlesworth, D., Zhang, Y., Bergero, R., Graham, C., Gardner, J. and Yong, L.

(2020) Using GC content to compare recombination patterns on the sex chromosomes and autosomes of the guppy, and its close outgroup species, *Molecular Biology and Evolution*, pp. 1–36.
Chevin, L. M. and Hospital, F. (2008) Selective sweep at a quantitative trait locus in the presence of background genetic variation, *Genetics*, 180, pp. 1645–1660. doi: 10.1534/genetics.108.093351.
Christin, P. A., Salamin, N., Savolainen, V., Duvall, M. R. and Besnard, G. (2007) C4 photosynthesis evolved in grasses via parallel adaptive genetic changes, *Current Biology*, 17, pp. 1241–1247. doi: 10.1016/j.cub.2007.06.036.
Collins, S. (2011) Many possible worlds: Expanding the ecological scenarios in experimental evolution, *Evolutionary Biology*, 38, pp. 3–14. doi: 10.1007/s11692-010-9106-3.
Colosimo, P. F. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles, *Science*, 307, pp. 1928–1933. doi: 10.1126/science.1107239.
Constanz, G. D. (1989) Reproductive biology of poeciliid fishes, in Meffe, G. K. and Snelson, F. F. (eds) *Ecology and evolution of livebearing fishes (Poecilidae)*. Englewood Cliffs, NJ: Prentice Hall, pp. 33–50.
Conte, G. L., Arnegard, M. E., Peichel, C. L. and Schluter, D. (2012) The probability of genetic parallelism and convergence in natural populations, *Proceedings of the Royal Society B: Biological Sciences*, 279, pp. 5039–5047. doi: 10.1098/rspb.2012.2146.
Cook, B. D., Kennard, M. J., Real, K., Pusey, B. J. and Hughes, J. M. (2011) Landscape genetic analysis of the tropical freshwater fish Mogurnda mogurnda (Eleotridae) in a monsoonal river basin: Importance of hydrographic factors and population history, *Freshwater Biology*, 56, pp. 812–827. doi: 10.1111/j.1365-2427.2010.02527.x.
Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W. and Pritchard, J. K. (2009) The role of geography in human adaptation, *PLoS Genetics*, 5(6), p. e1000500. doi: 10.1371/journal.pgen.1000500.
Crispo, E., Bentzen, P., Reznick, D. N., Kinnison, M. T. and Hendry, A. P. (2006) The relative influence of natural selection and geography on gene flow in guppies, *Molecular Ecology*, 15, pp. 49–62. doi: 10.1111/j.1365-294X.2005.02764.x.
Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G. and Durbin, R. (2011) The variant call format and VCFtools, *Bioinformatics*, 27(15), pp. 2156–2158. doi: 10.1093/bioinformatics/btr330.
Daniel, M. J. and Rodd, F. H. (2016) Female guppies can recognize kin but only avoid incest when previously mated, *Behavioral Ecology*, 27(1), pp. 55–62.
Deacon, A. E., Jones, F. A. M. and Magurran, A. E. (2018) Gradients in predation risk in a tropical river system, *Current Zoology*, 64(2), pp. 213–221. doi: 10.1093/cz/zoy004.
Delaneau, O., Marchini, J. and Zagury, J. F. (2012) A linear complexity phasing method for thousands of genomes, *Nature Methods*, 9(2), pp. 179–181. doi: 10.1038/nmeth.1785.
Depaulis, F., Mousset, S. and Veuille, M. (2003) Power of neutrality tests to detect bottlenecks and hitchhiking, *Journal of Molecular Evolution*, 57, pp. 190–200. doi: 10.1007/s00239-003-0027-y.
Dlugosch, K. M. and Parker, I. M. (2008) Founding events in species invasions:

Genetic variation, adaptive evolution, and the role of multiple introductions, *Molecular Ecology*, 17, pp. 431–449. doi: 10.1111/j.1365-294X.2007.03538.x.
Dobzhansky, T. (1937) *Genetics and the Origin of Species.* 2nd edn, *American Midland Naturalist*. 2nd edn. Edited by L. C. Dunn. New York: Columbia University Press. doi: 10.2307/2420611.
Dubois, Q., Lebigre, C., Schtickzelle, N. and Turlure, C. (2018) Sex, size and timing: Sampling design for reliable population genetics analyses using microsatellite data, *Methods in Ecology and Evolution*, 9(4), pp. 1036–1048. doi: 10.1111/2041-210X.12948.
El-Sabaawi, R. W., Frauendorf, T. C., Marques, P. S., Mackenzie, R. A., Manna, L. R., Mazzoni, R., Phillip, D. A. T., Warbanski, M. L. and Zandonà, E. (2016) Biodiversity and ecosystem risks arising from using guppies to control mosquitoes, *Biology letters*, 12(10). doi: 10.1098/rsbl.2016.0590.
Elleouet, J. S. and Aitken, S. N. (2018) Exploring Approximate Bayesian Computation for inferring recent demographic history with genomic markers in nonmodel species, *Molecular Ecology Resources*, 18, pp. 525–540. doi: 10.1111/1755-0998.12758.
Endler, J. A. (1980) Natural selection on color patterns in Poecilia reticulata, *Evolution*, 34(1), pp. 76–91. Available at: http://www.mhhe.com/biosci/genbio/raven6/lab6/labs/lab6/resources/original.pdf (Accessed: 3 May 2017).
Endler, J. A. and Houde, A. E. (1995) Geographic variation in female preferences for male traits in Poecilia reticulata, *Evolution*, 49(3), pp. 456–468. doi: 10.2307/2410270.
Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. and Foll, M. (2013) Robust demographic inference from genomic and SNP data, *PLoS Genetics*, 9(10), p. e1003905. doi: 10.1371/journal.pgen.1003905.
Falconer, D. S. and Mackay, T. F. C. (1996) *Introduction to quantitative genetics*. Fourth edi. Essex, England: Longman.
Fay, J. C. and Wu, C. I. (2000) Hitchhiking under positive Darwinian selection, *Genetics*, 155(3), pp. 1405–1413. Available at: http://www.genetics.org/content/genetics/155/3/1405.full.pdf (Accessed: 12 June 2017).
Fitzpatrick, J. L. and Evans, J. P. (2014) Postcopulatory inbreeding avoidance in guppies, *Journal of Evolutionary Biology*, 27(12), pp. 2585–2594. doi: 10.1111/jeb.12545.
Fitzpatrick, S. W., Gerberich, J. C., Kronenberger, J. A., Angeloni, L. M. and Funk, W. C. (2015) Locally adapted traits maintained in the face of high gene flow, *Ecology Letters*, 18(1), pp. 37–47. doi: 10.1111/ele.12388.
Fitzpatrick, S. W., Torres-Dowdall, J., Reznick, D. N., Ghalambor, C. K. and Funk, W. C. (2014) Parallelism isn't perfect: Could disease and flooding drive a life-history anomaly in Trinidadian guppies?, *American Naturalist*, 183(2), pp. 290–300. doi: 10.1086/674611.
Fraser, B. A., Künstner, A., Reznick, D. N., Dreyer, C. and Weigel, D. (2015) Population genomics of natural and experimental populations of guppies (Poecilia reticulata), *Molecular Ecology*, 24(2), pp. 389–408. doi: 10.1111/mec.13022.
Fraser, B. A. and Whiting, J. R. (2019) What can be learned by scanning the genome for molecular convergence in wild populations?, *Annals of the New York Academy of Sciences*, pp. 1–20. doi: 10.1111/nyas.14177.
Fraser, B. A., Whiting, J. R., Paris, J. R., Weadick, C. J., Parsons, P. J.,

Charlesworth, D., Bergero, R., Bemm, F., Hoffmann, M., Kottler, V. A., Liu, C., Dreyer, C. and Weigel, D. (2020) Improved reference genome uncovers novel sex-linked regions in the guppy (, *Genome Biology and Evolution*, pp. 1–38. doi: https://doi.org/10.1093/gbe/evaa187.

Fu, Y. X. and Li, W. H. (1993) Statistical tests of neutrality of mutations, *Genetics*, 133, pp. 693–709. Available at: http://www.genetics.org/content/genetics/133/3/693.full.pdf.

Fukuzawa, T. and Obika, M. (1995) N-CAM and N-Cadherin are specifically axpressed in xanthophores, but not in the other types of pigment cells, melanophores, and iridiphores, *Pigment Cell Research*, 8(1), pp. 1–9. doi: 10.1111/j.1600-0749.1995.tb00768.x.

Galtier, N., Depaulis, F. and Barton, N. H. (2000) Detecting bottlenecks and selective sweeps from DNA sequence polymorphism, *Genetics*, 155(2), pp. 981–987. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9725864 (Accessed: 14 March 2019).

Garant, D., Forde, S. E. and Hendry, A. P. (2007) The multifarious effects of dispersal and gene flow on contemporary adaptation, *Functional Ecology*, 21(3), pp. 434–443. doi: 10.1111/j.1365-2435.2006.01228.x.

Garland Jr., T. and Rose, M. R. (eds) (2009) *Experimental Evolution*. First edit. London: University of California Press. doi: 10.1017/CBO9781107415324.004.

Garud, N. R., Messer, P. W., Buzbas, E. O. and Petrov, D. A. (2015) Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps, *PLoS Genetics*, 11(2), p. e1005004. doi: 10.1371/journal.pgen.1005004.

Gautier, M. (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates, *Genetics*, 201(4), pp. 1555–1579. doi: 10.1534/genetics.115.181453.

Gillespie, J. H. (1991) *The causes of molecular evolution*. Oxford: Oxford University Press. Available at: https://tinyurl.com/y4xk5r6a.

Gompert, Z. (2020) A population-genomic approach for estimating selection on polygenic traits in heterogeneous environments, *bioRxiv*, pp. 1–44. doi: 10.1101/2020.06.02.129874.

Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. and Desai, M. M. (2017) The dynamics of molecular evolution over 60,000 generations, *Nature*, 551, pp. 45–50. doi: 10.1038/nature24287.

Gordon, S. P., López-Sepulcre, A., Rumbo, D. and Reznick, D. N. (2016) Rapid changes in the sex linkage of male coloration in Introduced guppy populations, *The American Naturalist*, 189(2), pp. 196–200. doi: 10.1086/689864.

Grant, P. R. and Grant, B. R. (2008) *How and why species multiply: The radiation of Darwin's finches*. Princeton, NJ: Princeton University Press. Available at: https://books.google.co.uk/books?hl=en&lr=&id=h5urQiWUB7AC&oi=fnd&pg=PR1&ots=2t4W0yG0Pf&sig=wvUIT-4-CP-HQGqZjg8jxNzb0QM&redir_esc=y#v=onepage&q&f=false.

Graves Jr, J. L., Hertweck, K. L., Phillips, M. A., Han, M. V., Cabral, L. G., Barter, T. T., Greer, L. F., Burke, M. K., Mueller, L. D. and Rose, M. R. (2017) Genomics of parallel experimental evolution in drosophila, *Molecular Biology and Evolution*, 34(4), pp. 831–842. doi: 10.1093/molbev/msw282.

Green, D. A. and Extavour, C. G. (2012) Convergent evolution of a reproductive trait through distinct developmental mechanisms in Drosophila,

Developmental Biology. doi: 10.1016/j.ydbio.2012.09.014.
Grether, G. F., Millie, D. F., Bryant, M. J., Reznick, D. N. and Mayea, W. (2001) Rain forest canopy cover, resource availablity, and life history evolution in guppies, *Ecology*, 82(6), pp. 1546–1559. doi: 10.1890/0012-9658(2001)082[1546:RFCCRA]2.0.CO;2.
Guy, T. J., Gresswell, R. E. and Banks, M. A. (2008) Landscape-scale evaluation of genetic structure among barrier-isolated populations of coastal cutthroat trout, Oncorhynchus clarkii clarkii, *Canadian Journal of Fisheries and Aquatic Sciences*, 65(8), pp. 1749–1762. doi: 10.1139/F08-090.
Hamilton, J. A. and Miller, J. M. (2016) Adaptive introgression as a resource for management and genetic conservation in a changing climate, *Conservation Biology*, 30(1), pp. 33–41. doi: 10.1111/cobi.12574.
Hamilton, K. E., Ferguson, A. and Taggart, J. B. (1989) Ldh-5 as a phylogeographic marker locus, *Journal of Fish Biology*, 35, pp. 651–664.
Handelsman, C. A., Broder, E. D., Dalton, C. M., Ruell, E. W., Myrick, C. A., Reznick, D. N. and Ghalambor, C. K. (2013) Predator-induced phenotypic plasticity in metabolism and rate of growth: Rapid adaptation to a novel environment, *Integrative and Comparative Biology*, 53(6), pp. 975–988. doi: 10.1093/icb/ict057.
Harris, K. and Nielsen, R. (2013) Inferring demographic history from a spectrum of shared haplotype lengths, *PLoS Genetics*, 9(6), p. e1003521. doi: 10.1371/journal.pgen.1003521.
Haskins, C. P., Haskins, E. F., McLaughlin, J. J. and Hewitt, R. E. (1961) Polymorphism and population structure in Lebistes reticulatus, a population study, in Blair, W. F. (ed.) *Vertebrate Speciation*. Austin: University of Texas Press, pp. 320–395.
La Haye, M. J. J., Reiners, T. E., Raedts, R., Verbist, V. and Koelewijn, H. P. (2017) Genetic monitoring to evaluate reintroduction attempts of a highly endangered rodent, *Conservation Genetics*, 18(4), pp. 877–892. doi: 10.1007/s10592-017-0940-z.
Hermisson, J. and Pennings, P. S. (2005) Soft sweeps: Molecular population genetics of adaptation from standing genetic variation, *Genetics*, 169(4), pp. 2335–2352. doi: 10.1534/genetics.104.036947.
Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A. and Whitlock, M. C. (2016) Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions, *The American Naturalist*, 188(4), pp. 379–397. doi: 10.1086/688018.
Hoekstra, H. E. and Nachman, M. W. (2003) Different genes underlie adaptive melanism in different populations of rock pocket mice, *Molecular Ecology*, 12(5), pp. 1185–1194. doi: 10.1046/j.1365-294X.2003.01788.x.
Hooper, R., Excoffier, L., Forney, K. A., Gilbert, M. T. P., Martin, M. D., Morin, P. A., Wolf, J. B. W. and Foote, A. D. (2020) Runs of homozygosity in killer whale genomes provide a global record of demographic histories, *bioRxiv*. doi: 10.1017/CBO9781107415324.004.
Hudson, R. R., Slatkin, M. and Maddison, W. P. (1992) Estimation of levels of gene flow from DNA sequence data, *Genetics*, 132(2), pp. 583–589.
Hughes, J. M., Bunn, S. E., Kingston, D. M. and Hurwood, D. A. (1995) Genetic differentiation and dispersal among populations of Paratya australiensis (Atyidae) in rainforest streams in southeast Queensland, Australia, *Journal of the North American Benthological Society*, 14(1), pp. 158–173. doi:

10.2307/1467731.
Hughes, K. A., Houde, A. E., Price, A. C. and Rodd, F. H. (2013) Mating advantage for rare males in wild guppy populations, *Nature*, 503, pp. 108–110. doi: 10.1038/nature12717.
Jacobs, A., Carruthers, M., Yurchenko, A., Gordeeva, N. V., Alekseyev, S. S., Hooker, O., Leong, J. S., Minkley, D. R., Rondeau, E. B., Koop, B. F., Adams, C. E. and Elmer, K. R. (2020) Parallelism in eco-morphology and gene expression despite variable evolutionary and genomic backgrounds in a Holarctic fish, *PLoS Genetics*, 16(4), p. e1008658. doi: 10.1371/journal.pgen.1008658.
Jain, K. and Stephan, W. (2015) Response of polygenic traits under stabilizing selection and mutation when loci have unequal effects, *G3: Genes, Genomes, Genetics*, 5(6), pp. 1065–1074. doi: 10.1534/g3.115.017970.
Jain, K. and Stephan, W. (2017) Rapid adaptation of a polygenic trait after a sudden environmental shift, *Genetics*, 206(1), pp. 389–406. doi: 10.1534/genetics.116.196972.
Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks, *Nature*, 484, pp. 55–61. doi: 10.1038/nature10944.
Kang, L., Rashkovetsky, E., Michalak, K., Garner, H. R., Mahaney, J. E., Rzigalinski, B. A., Korol, A., Nevo, E. and Michalak, P. (2019) Genomic divergence and adaptive convergence in Drosophila simulans from Evolution Canyon, Israel, *Proceedings of the National Academy of Sciences of the United States of America*, 116(24), pp. 11839–11844. doi: 10.1073/pnas.1720938116.
Kardos, M., Qvarnström, A. and Ellegren, H. (2017) Inferring individual inbreeding and demographic history from segments of identity by descent in Ficedula flycatcher genome sequences, *Genetics*, 205(3), pp. 1319–1334. doi: 10.1534/genetics.116.198861.
Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I. and Whitlock, M. C. (2012) Experimental evolution, *Trends in Ecology and Evolution*, 27(10), pp. 547–560. doi: 10.1016/j.tree.2012.06.001.
Kelley, J. L. and Magurran, A. E. (2003) Effects of relaxed predation pressure on visual predator recognition in the guppy, *Behavioral Ecology and Sociobiology*, 54, pp. 225–232. doi: 10.1007/s00265-003-0621-4.
Kemp, D. J., Batistic, F.-K. and Reznick, D. N. (2018) Predictable adaptive trajectories of sexual coloration in the wild: evidence from replicate experimental guppy populations, *Evolution*, 72(11), pp. 2462–2477. doi: 10.1111/evo.13564.
Kemp, D. J., Reznick, D. N., Grether, G. F. and Endler, J. A. (2009) Predicting the direction of ornament evolution in Trinidadian guppies (Poecilia reticulata)., *Proceedings of the Royal Society B: Biological Sciences*, 276(1677), pp. 4335–4343. doi: 10.1098/rspb.2009.1226.
Kim, Y. and Nielsen, R. (2004) Linkage disequilibrium as a signature of selective sweeps, *Genetics*, 167(3), pp. 1513–1524. doi: 10.1534/genetics.103.025387.
Kingman, J. F. C. (1982) The coalescent, *Stochastic Processes and their Applications*, 13(3), pp. 235–248. doi: 10.1016/0304-4149(82)90011-4.
Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., Mckeigue, P. M. and

Wilson, J. F. (2010) Genomic runs of homozygosity record population history and consanguinity, *PLoS ONE*, 5(11), p. e13996. doi: 10.1371/journal.pone.0013996.
Kohler, T. J., Heatherly, T. N., El-Sabaawi, R. W., Zandonà, E., Marshall, M. C., Flecker, A. S., Pringle, C. M., Reznick, D. N. and Thomas, S. A. (2012) Flow, nutrients, and light availability influence Neotropical epilithon biomass and stoichiometry, *Freshwater Science*, 31(4), pp. 1019–1034. doi: 10.1899/11-141.1.
Korneliussen, T. S., Albrechtsen, A. and Nielsen, R. (2014) ANGSD: Analysis of Next Generation Sequencing Data, *BMC Bioinformatics*, 15, p. 356. doi: 10.1186/s12859-014-0356-4.
Krueger, F. (2012) TrimGalore! Available at: https://github.com/FelixKrueger/TrimGalore.
Künstner, A., Hoffmann, M., Fraser, B. A., Kottler, V. A., Sharma, E., Weigel, D. and Dreyer, C. (2016) The genome of the Trinidadian guppy, Poecilia reticulata , and variation in the Guanapo population, *PloS one*, 11(12), pp. 1–25. doi: 10.1371/journal.pone.0169087.
Lachapelle, J., Reid, J. and Colegrave, N. (2015) Repeatability of adaptation in experimental populations of different sizes, *Proceedings of the Royal Society B: Biological Sciences*, 282, p. 20143033. doi: 10.1098/rspb.2014.3033.
Lambert, D. M., King, T., Shepherd, L. D., Livingston, A., Anderson, S. and Craig, J. L. (2005) Serial population bottlenecks and genetic variation: Translocated populations of the New Zealand Saddleback (Philesturnus carunculatus rufusater), *Conservation Genetics*, 6, pp. 1–14. doi: 10.1007/s10592-004-7857-z.
Lande, R. (2009) Adaptation to an extraordinary environment by evolution of phenotypic plasticity and genetic assimilation, *Journal of Evolutionary Biology*, 22, pp. 1435–1446. doi: 10.1111/j.1420-9101.2009.01754.x.
Lavergne, S., Mouquet, N., Thuiller, W. and Ronce, O. (2010) Biodiversity and climate change: Integrating evolutionary and ecological responses of species and communities, *Annual Review of Ecology, Evolution, and Systematics*, 41, pp. 321–350. doi: 10.1146/annurev-ecolsys-102209-144628.
Lee, K. M. and Coop, G. (2019) Population genomics perspectives on convergent adaptation, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1777), p. 20180236. doi: 10.1098/rstb.2018.0236.
Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.
Li, Y., Cheng, R., Spokas, K. A., Palmer, A. A. and Borevitz, J. O. (2014) Genetic variation for life history sensitivity to seasonal warming in Arabidopsis thaliana, *Genetics*, 196(2), pp. 569–577. doi: 10.1534/genetics.113.157628.
Liao, H., McKenzie, T. and Hageman, R. (1986) Isolation of a thermostable enzyme variant by cloning and selection in a thermophile, *Proceedings of the National Academy of Sciences of the United States of America*, 83(3), pp. 576–580. doi: 10.1073/pnas.83.3.576.
De Lisle, S. P. and Bolnick, D. I. (2020) A multivariate view of parallel evolution, *Evolution*, 74(7), pp. 1466–1481. doi: 10.1111/evo.14035.
Lopez-Sepulcre, A., Gordon, S. P., Paterson, I. G., Bentzen, P. and Reznick, D. N. (2013) Beyond lifetime reproductive success: the posthumous

reproductive dynamics of male Trinidadian guppies, *Proceedings of the Royal Society B: Biological Sciences*, 280(1763), pp. 20131116–20131116. doi: 10.1098/rspb.2013.1116.
Losos, J. B. (2009) *Lizards in an evolutionary tree: ecology and adaptive radiation of Anoles*. First edit. Oakland, California: University of California Press.
Losos, J. B., Warheitt, K. I. and Schoener, T. W. (1997) Adaptive differentiation following experimental island colonization in Anolis lizards, *Nature*, 387, pp. 70–73. doi: 10.1038/387070a0.
Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F. and Storfer, A. (2017) Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation, *Molecular ecology resources*, 17, pp. 142–152. doi: 10.1111/1755-0998.12635.
MacPherson, A. and Nuismer, S. L. (2017) The probability of parallel genetic evolution from standing genetic variation, *Journal of Evolutionary Biology*, 30(2), pp. 326–337. doi: 10.1111/jeb.13006.
Magurran, A. E. (2005) *Evolutionary Ecology: Trinidadian Guppy*. First edit, *Oxford University Press*. First edit. Oxford University Press. doi: 10.1093/acprof.
Magurran, A. E., Seghers, B. H., Carvalho, G. R. and Shaw, P. W. (1992) Behavioural consequences of an artificial introduction of guppies (Poecilia reticulata) in N. Trinidad: evidence for the evolution of anti-predator behaviour in the wild, *Proceedings of the Royal Society of London B: Biological Sciences*, 248, pp. 117–122.
Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C. and Sedlazeck, F. J. (2019) Structural variant calling: the long and the short of it, *Genome Biology*, 20. doi: 10.1186/s13059-019-1828-7.
Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F. and Durbin, R. (2018) Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow, *Nature Ecology & Evolution*, 2, pp. 1940–1955. doi: 10.1038/s41559-018-0717-x.
Marques, D. A., Jones, F. C., Di Palma, F., Kingsley, D. M. and Reimchen, T. E. (2018) Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation, *Nature Ecology and Evolution*, 2, pp. 1128–1138. doi: 10.1038/s41559-018-0581-8.
Martin, S. H., Davey, J. W., Salazar, C. and Jiggins, C. D. (2019) Recombination rate variation shapes barriers to introgression across butterfly genomes, *PLoS biology*, 17(2). doi: 10.1101/297531.
Matos, M., Simões, P., Santos, M. A., Seabra, S. G., Faria, G. S., Vala, F., Santos, J. and Fragata, I. (2015) History, chance and selection during phenotypic and genomic experimental evolution: Replaying the tape of life at different levels, *Frontiers in Genetics*, 6, p. 71. doi: 10.3389/fgene.2015.00071.
Mayr, E. (1963) *Animal species and evolution*. Cambridge: Harvard University Press.
Meier, J. I., Marques, D. A., Wagner, C. E., Excoffier, L. and Seehausen, O. (2018) Genomics of parallel ecological speciation in Lake Victoria cichlids, *Molecular Biology and Evolution*, 35(6), pp. 1489–1506. doi: 10.1093/molbev/msy051.
Le Moan, A., Bekkevold, D. and Hemmer-Hansen, J. (2020) Evolution at two time-frames: ancient and common origin of two structural variants involved

in local adaptation of the European plaice (Pleuronectes platessa), *bioRxiv*. doi: 10.1101/662577.
Nachman, M. W., Hoekstra, H. E. and D'Agostino, S. L. (2003) The genetic basis of adaptive melanism in pocket mice, *Proceedings of the National Academy of Sciences of the United States of America*, 100(9), pp. 5268–5273. doi: 10.1073/pnas.0431157100.
Nei, M. (1987) *Molecular Evolutionary Genetics*. New York: Columbia University Press. doi: 10.7312/nei-92038.
Nei, M. and Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases, *Proceedings of the National Academy of Sciences of the United States of America*, 76(10), pp. 5269–5273. doi: 10.1073/pnas.76.10.5269.
Nei, M., Maruyama, T. and Chakraborty, R. (1975) The bottleneck effect and genetic variability in populations, *The Society for the Study of Evolution*, 29(1), pp. 1–10. doi: 10.1111/j.1558-5646.1975.tb00807.x.
Nelson, W. J. and Nusse, R. (2004) Convergence of Wnt, β-catenin, and cadherin pathways, *Science*, 303, pp. 1483–1487. doi: 10.1126/science.1094291.
Nielsen, R. (2005) Molecular signatures of natural selection, *Annual Review of Genetics*, 39, pp. 197–218. doi: 10.1146/annurev.genet.39.073003.112420.
Nishimura, E. K., Yoshida, H., Kunisada, T. and Nishikawa, S. I. (1999) Regulation of E- and P-cadherin expression correlated with melanocyte migration and diversification, *Developmental Biology*, 215(2), pp. 155–166. doi: 10.1006/dbio.1999.9478.
Orr, H. A. (2005) The genetic theory of adaptation: a brief history, *Nature Reviews Genetics*, 6, pp. 119–127. doi: 10.1038/nrg1523.
Ortego, J., Aparicio, J. M., Calabuig, G. and Cordero, P. J. (2007) Increase of heterozygosity in a growing population of lesser kestrels, *Biology Letters*, 3, pp. 585–588. doi: 10.1098/rsbl.2007.0268.
Overcast, I. (2017) EasySFS. Available at: https://github.com/isaacovercast/easySFS.
Pannell, J. R. (2003) Coalescence in a metapopulation with recurrent local extinction and recolonization, *Evolution*, 57(5), pp. 949–961. doi: 10.1111/j.0014-3820.2003.tb00307.x.
Pfeifer, B. (2014) Whole genome analyses using PopGenome and VCF files, pp. 1–28. Available at: https://cran.r-project.org/web/packages/PopGenome/vignettes/Whole_genome_analyses_using_VCF_files.pdf (Accessed: 27 April 2017).
Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. and Lercher, M. J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R, *Molecular Biology and Evolution*, 31(7), pp. 1929–1936. doi: 10.1093/molbev/msu136.
Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., *et al.* (2013) Great ape genetic diversity and population history, *Nature*, 499, pp. 471–475. doi: 10.1038/nature12228.
Price, T. D., Qvarnström, A. and Irwin, D. E. (2003) The role of phenotypic plasticity in driving genetic evolution, *Proceedings of the Royal Society B: Biological Sciences*, 270, pp. 1433–1440. doi: 10.1098/rspb.2003.2372.
Pritchard, J. K., Pickrell, J. K. and Coop, G. (2010) The genetics of human

adaptation: hard sweeps, soft sweeps, and polygenic adaptation, *Current Biology*, 20(4), pp. R208–R215. doi: 10.1038/jid.2014.371.
Pritchard, J. K. and Di Rienzo, A. (2010) Adaptation - Not by sweeps alone, *Nature Reviews Genetics*, 11(10), pp. 665–667. doi: 10.1038/nrg2880.
Przeworski, M. (2002) The signature of positive selection at randomly chosen loci, *Genetics*, 160(3), pp. 1179–1189.
Purcell, J., Brelsford, A., Wurm, Y., Perrin, N. and Chapuisat, M. (2014) Convergent genetic architecture underlies social organization in ants, *Current Biology*, 24(22), pp. 2728–2732. doi: 10.1016/j.cub.2014.09.071.
Purfield, D. C., McParland, S., Wall, E. and Berry, D. P. (2017) The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds, *PLoS ONE*, 12(5), p. e0176780. doi: 10.1371/journal.pone.0176780.
Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I. and Bird, C. E. (2014) Demystifying the RAD fad, *Molecular Ecology*, 23, pp. 5937–5942. doi: 10.1111/mec.12965.
Raposo do Amaral, F., Maldonado-Coelho, M., Aleixo, A., Luna, L. W., Rêgo, P. S. do, Araripe, J., Souza, T. O., Silva, W. A. G. and Thom, G. (2018) Recent chapters of Neotropical history overlooked in phylogeography: shallow divergence explains phenotype and genotype uncoupling in Antilophia manakins, *Molecular Ecology*, 27, pp. 4108–4120. doi: 10.1111/mec.14843.
Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., Bowers, J. E., Burke, J. M. and Rieseberg, L. H. (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers, *Nature Communications*, 4, p. 1827. doi: 10.1038/ncomms2833.
Rennison, D. J., Rudman, S. M. and Schluter, D. (2019) Genetics of adaptation: Experimental test of a biotic mechanism driving divergence in traits and genes, *Evolution Letters*, 3, pp. 513–520. doi: 10.1002/evl3.135.
Reznick, D. N. (1982) The impact of predation on life history evolution in Trinidadian guppies: genetic basis of observed life history patterns, *Evolution*, 36(6), p. 1236. doi: 10.2307/2408156.
Reznick, D. N. (1989) Life-history evolution in guppies: 2. Repeatability of field observations and the effects of season on life histories, *Evolution*, 43(6), pp. 1285–1297. doi: 10.2307/2409363.
Reznick, D. N., Bassar, R. D., Handelsman, C. A., Ghalambor, C. K., Arendt, J. D., Coulson, T., Potter, T., Ruell, E. W., Torres-Dowdall, J., Bentzen, P. and Travis, J. (2019) Eco-evolutionary feedbacks predict the time course of rapid life history evolution, *The American Naturalist*, 194(5), pp. 671–692. doi: 10.1086/705380.
Reznick, D. N. and Bryga, H. A. (1987) Life-history evolution in guppies (Poecilia reticulata): I. Phenotypic and genetic changes in an introduction experiment, *Evolution*, 41(6), pp. 1370–1385. Available at: http://www.jstor.org/stable/2409101 (Accessed: 3 May 2017).
Reznick, D. N. and Bryga, H. A. (1996) Life-history evolution in guppies (Poecilia reticulata: Poeciliidae). V. Genetic basis of parallelism in life histories, *American Naturalist*, 147(3), pp. 339–359. doi: 10.1086/285855.
Reznick, D. N., Bryga, H. A. and Endler, J. A. (1990) Experimentally induced life-history evolution in a natural population, *Nature*, 346, pp. 357–359. doi: 10.1038/346357a0.
Reznick, D. N., Butler IV, M. J. and Rodd, F. H. (2001) Life-history evolution in guppies. VII. The comparative ecology of high- and low-predation

environments, *American Naturalist*, 157, pp. 126–140. doi: 10.1086/318627.
Reznick, D. N. and Endler, J. A. (1982) The impact of predation on life history Evolution in Trinidadian guppies (Poecilia reticulata), *Evolution*, 36(1), pp. 160–177. doi: 10.2307/2408156.
Reznick, D. N., Losos, J. B. and Travis, J. (2019) From low to high gear: there has been a paradigm shift in our understanding of evolution, *Ecology Letters*, 22, pp. 233–244. doi: 10.1111/ele.13189.
Reznick, D. N., Shaw, F. H., Rodd, F. H. and Shaw, R. G. (1997) Evaluation of the rate of evolution in natural populations of guppies (Poecilia reticulata), *Science*, 275(5308), pp. 1934–1937. doi: 10.1126/science.275.5308.1934.
Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., *et al.* (2009) Planetary boundaries: Exploring the safe operating space for humanity, *Ecology and Society*, 14(2), p. 32. doi: 10.5751/ES-03180-140232.
Roelants, K., Fry, B. G., Norman, J. A., Clynen, E., Schoofs, L. and Bossuyt, F. (2010) Identical skin toxins by convergent molecular adaptation in frogs, *Current Biology*, 20(2), pp. 125–130. doi: 10.1016/J.CUB.2009.11.015.
Roman, J. and Darling, J. A. (2007) Paradox lost: genetic diversity and the success of aquatic invasions, *Trends in Ecology & Evolution*, 22(9), pp. 454–464. doi: 10.1016/J.TREE.2007.07.002.
Rosenblum, E. B., Parent, C. E. and Brandt, E. E. (2014) The molecular basis of phenotypic convergence, *Annual Review of Ecology, Evolution, and Systematics*, 45(1), pp. 203–226. doi: 10.1146/annurev-ecolsys-120213-091851.
Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. and Lander, E. S. (2002) Detecting recent positive selection in the human genome from haplotype structure, *Nature*, 419, pp. 832–837. doi: 10.1038/nature01140.
Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S. and Consortium, T. I. H. (2007) Genome-wide detection and characterization of positive selection in human populations, *Nature*, 449, pp. 913–918. doi: 10.1038/nature06250.
Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J. and Schluter, D. (2017) Gene flow and selection interact to promote adaptive divergence in regions of low recombination, *Molecular Ecology*, 26, pp. 4378–4390. doi: 10.1111/mec.14226.
Sane, M., Diwan, G. D., Bhat, B. A., Wahl, L. M. and Agashe, D. (2020) Shifts in mutation spectra enhance access to beneficial mutations, *bioRxiv*, pp. 1–21.
Santiago, E., Novo, I., Pardiñas, A., Saura, M., Wang, J. and Caballero, A. (2020) Recent demographic history inferred by high-resolution analysis of linkage disequilibrium, *Molecular Biology and Evolution*, pp. 1–29. doi: 10.1093/molbev/msaa169.
Sard, N., Robinson, J., Kanefsky, J., Herbst, S. and Scribner, K. (2019) Coalescent models characterize sources and demographic history of recent round goby colonization of Great Lakes and inland waters, *Evolutionary*

Applications, 12, pp. 1034–1049. doi: 10.1111/eva.12779.
Schluter, D. (2000) Ecological character displacement in adaptive radiation, *The American Naturalist*, 156(S4), pp. S4–S16. doi: 10.2307/3079223.
Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A. and Schatz, M. C. (2018) Accurate detection of complex structural variations using single-molecule sequencing, *Nature Methods*, 15, pp. 461–468. doi: 10.1038/s41592-018-0001-7.
Seghers, B. H. (1974) Schooling behavior in the guppy (Poecilia reticulata): an evolutionary response to predation, *Evolution*, 28(3), pp. 486–489.
Shaw, P. W., Carvalho, G. R., Magurran, A. E. and Seghers, B. H. (1991) Population differentiation in Trinidadian guppies (Poecilia reticulata): patterns and problems, *Journal of Fish Biology*, 39, pp. 203–209. doi: 10.1111/j.1095-8649.1991.tb05084.x.
Shaw, P. W., Carvalho, G. R., Seghers, B. H. and Magurran, A. E. (1992) Genetic consequences of an artificial introduction of guppies (Poecilia reticulata) in N. Trinidad, *Proceedings of the Royal Society of London B: Biological Sciences*, 248, pp. 111–116.
Simões, P., Fragata, I., Seabra, S. G., Faria, G. S., Santos, M. A., Rose, M. R., Santos, M. and Matos, M. (2017) Predictable phenotypic, but not karyotypic, evolution of populations with contrasting initial history, *Scientific Reports*, 7, pp. 1–12. doi: 10.1038/s41598-017-00968-1.
Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., Hooper, D. M., Strand, A. I., Li, Q., Raney, B., Balakrishnan, C. N., Griffith, S. C., Mcvean, G. and Przeworski, M. (2015) Stable recombination hotspots in birds, *Science*, 350(6263), pp. 928–932.
Smith, J. M. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene, *Genetical Research*, 23, pp. 23–35. doi: 10.1017/S0016672300014634.
Speechley, E. M., Gasparini, C. and Evans, J. P. (2019) Female guppies increase their propensity for polyandry as an inbreeding avoidance strategy, *Animal Behaviour*, 157, pp. 87–93. doi: 10.1016/j.anbehav.2019.08.016.
Stankowski, S., Chase, M. A., Fuiten, A. M., Rodrigues, M. F., Ralph, P. L. and Streisfeld, M. A. (2019) Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers, *PLoS Biology*, 17(7), p. e3000391. doi: 10.1371/journal.pbio.3000391.
Steiner, C. C., Römpler, H., Boettger, L. M., Schöneberg, T. and Hoekstra, H. E. (2009) The genetic basis of phenotypic convergence in beach mice: Similar pigment patterns but different genes, *Molecular Biology and Evolution*, 26(1), pp. 35–45. doi: 10.1093/molbev/msn218.
Stephan, W. (2016) Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation, *Molecular Ecology*, 25, pp. 79–88. doi: 10.1111/mec.13288.
Stern, D. L. (2013) The genetic causes of convergent evolution, *Nature Reviews Genetics*, 14, pp. 751–764. doi: 10.1038/nrg3483.
Stoltzfus, A. and McCandlish, D. M. (2017) Mutational biases influence parallel adaptation, *Molecular Biology and Evolution*, 34(9), pp. 2163–2172. doi: 10.1093/molbev/msx180.
Storz, J. F., Natarajan, C., Signore, A. V., Witt, C. C., McCandlish, D. M. and Stoltzfus, A. (2019) The role of mutation bias in adaptive molecular evolution: Insights from convergent changes in protein function, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1777), p. 20180238. doi: 10.1098/rstb.2018.0238.

Stuart, Y. E., Veen, T., Weber, J. N., Hanson, D., Ravinet, M., Lohman, B. K., Thompson, C. J., Tasneem, T., Doggett, A., Izen, R., Ahmed, N., Barrett, R. D. H., Hendry, A. P., Peichel, C. L. and Bolnick, D. I. (2017) Contrasting effects of environment and genetics generate a continuum of parallel evolution, *Nature Ecology and Evolution*, 1, p. 0158. doi: 10.1038/s41559-017-0158.

Suk, H. Y. and Neff, B. D. (2009) Microsatellite genetic differentiation among populations of the Trinidadian guppy, *Heredity*, 102, pp. 425–434. doi: 10.1038/hdy.2009.7.

Szpiech, Z. A. and Hernandez, R. D. (2014) Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection, *Molecular Biology and Evolution*, 31(10), pp. 2824–2827. doi: 10.1093/molbev/msu211.

Tajima, F. (1989a) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, *Genetics*, 123, pp. 585–595. Available at: http://www.genetics.org/content/genetics/123/3/585.full.pdf (Accessed: 12 June 2017).

Tajima, F. (1989b) The effect of change in population size on DNA polymorphism., *Genetics*, 123(3), pp. 597–601.

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Médigue, C., Schneider, D. and Lenski, R. E. (2016) Tempo and mode of genome evolution in a 50,000-generation experiment, *Nature*, 536, pp. 165–170. doi: 10.1038/nature18959.

Therkildsen, N. O., Wilder, A. P., Conover, D. O., Munch, S. B., Baumann, H. and Palumbi, S. R. (2019) Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing, *Science*, 365(6452), pp. 487–490. doi: 10.1126/science.aaw7271.

Thompson, K. A., Osmond, M. M. and Schluter, D. (2019) Parallel genetic evolution and speciation from standing variation, *Evolution Letters*, 3(2), pp. 129–141. doi: 10.1002/evl3.106.

Thorogood, C. J., Bauer, U. and Hiscock, S. J. (2018) Convergent and divergent evolution in carnivorous pitcher plant traps, *New Phytologist*, 217, pp. 1035–1041. doi: 10.1111/nph.14879.

Tian, C., Li, L., Liang, X. F., He, S., Guo, W., Lv, L., Wang, Q. and Song, Y. (2016) Identification of differentially expressed genes associated with differential body size in mandarin fish (Siniperca chuatsi), *Genetica*, 144, pp. 445–455. doi: 10.1007/s10709-016-9913-2.

Tiffin, P. and Ross-Ibarra, J. (2014) Advances and limits of using population genetics to understand local adaptation, *Trends in Ecology and Evolution*, 29(12), pp. 673–680. doi: 10.1016/j.tree.2014.10.004.

Torres-Dowdall, J., Handelsman, C. A., Reznick, D. N. and Ghalambor, C. K. (2012) Local adaptation and the evolution of phenotypic plasticity in Trinidadian guppies (Poecilia reticulata), *Evolution*, 66(11), pp. 3432–3443. doi: 10.1111/j.1558-5646.2012.01694.x.

Torres-Dowdall, J., Handelsman, C. A., Ruell, E. W., Auer, S. K., Reznick, D. N. and Ghalambor, C. K. (2012) Fine-scale local adaptation in life histories along a continuous environmental gradient in Trinidadian guppies, *Functional Ecology*, 26(3), pp. 616–627. doi: 10.1111/j.1365-2435.2012.01980.x.

Torres, R., Szpiech, Z. A. and Hernandez, R. D. (2018) Human demographic history has amplified the effects of background selection across the

genome, *PLoS Genetics*. Edited by G. Coop, 14(6), p. e1007387. doi: 10.1371/journal.pgen.1007387.
Tripathi, N., Hoffmann, M., Willing, E.-M., Lanz, C., Weigel, D. and Dreyer, C. (2009) Genetic linkage map of the guppy, Poecilia reticulata, and quantitative trait loci analysis of male size and colour variation., *Proceedings of the Royal Society B: Biological Sciences*, 276, pp. 2195–2208. doi: 10.1098/rspb.2008.1930.
Tse, W. K. F. (2017) Importance of deubiquitinases in zebrafish craniofacial development, *Biochemical and Biophysical Research Communications*, 487(4), pp. 813–819. doi: 10.1016/j.bbrc.2017.04.132.
van't Hof, A. E., Edmonds, N., Dalíková, M., Marec, F. and Saccheri, I. J. (2011) Industrial melanism in British peppered moths has a singular and recent mutational origin., *Science*, 332(6032), pp. 958–60. doi: 10.1126/science.1203043.
Venkatesh, B., Tan, C. H. and Lam, T. J. (1992) Prostaglandins and teleost neurohypophyseal hormones induce premature parturition in the guppy, Poecilia reticulata, *General and Comparative Endocrinology*, 87, pp. 28–32. doi: 10.1016/0016-6480(92)90146-B.
Vogwill, T., Phillips, R. L., Gifford, D. R. and Maclean, R. C. (2016) Divergent evolution peaks under intermediate population bottlenecks during bacterial experimental evolution, *Proceedings of the Royal Society B: Biological Sciences*, 283, p. 20160749. doi: 10.1098/rspb.2016.0749.
de Vries, C. and Caswell, H. (2019) Selection in two-sex stage-structured populations: Genetics, demography, and polymorphism, *Theoretical Population Biology*, 130, pp. 160–169. doi: 10.1016/j.tpb.2019.07.012.
Waddington, C. H. (1953) Genetic assimilation of an acquired character, *Evolution*, 7(2), pp. 118–126.
Wakeley, J. and Aliacar, N. (2001) Gene genealogies in a metapopulation, *Genetics*, 159, pp. 893–905.
Wang, H. X., Tekpetey, F. R. and Kidder, G. M. (2009) Identification of WNT/β-CATENIN signaling pathway components in human cumulus cells, *Molecular Human Reproduction*, 15(1), pp. 11–17. doi: 10.1093/molehr/gan070.
Wang, M., Zhao, Y. and Zhang, B. (2015) Efficient test and visualization of multi-set intersections, *Scientific Reports*, 5, p. 16923. doi: 10.1038/srep16923.
Wares, J. P., Hughes, a R. and Grosberg, R. K. (2005) Mechanisms that drive evolutionary change: insights from species introductions and invasions, in Sax, D. F., Stachowicz, J. J., and Gaines, S. D. (eds) *Species Invasions: Insights into Ecology, Evolution, and Biogeography*. Sunderland: Sinauer Associates Incorporated, pp. 229–257.
Weigand, H. and Leese, F. (2018) Detecting signatures of positive selection in non-model species using genomic data, *Zoological Journal of the Linnean Society*, 184, pp. 528–583. doi: 10.1093/zoolinnean/zly007.
Wein, T. and Dagan, T. (2019) The effect of population bottleneck size and selective regime on genetic diversity and evolvability in bacteria, *Genome Biology and Evolution*, 11(11), pp. 3283–3290. doi: 10.1093/gbe/evz243.
Weir, B. S. and Cockerham, C. C. (1984) Estimating F-Statistics for the Analysis of Population Structure, *Evolution*, 38(386), pp. 1358–1370. Available at: http://www.jstor.org/stable/2408641 (Accessed: 22 August 2017).
Whiting, J. R. and Fraser, B. A. (2020) Contingent convergence: The ability to detect convergent genomic evolution is dependent on population size and

migration, *G3: Genes, Genomes, Genetics*, 10(2), pp. 677–693. doi: 10.1534/g3.119.400970.
Whiting, J. R., Paris, J. R., Van Der Zee, M. J., Parsons, P. J. and Fraser, B. A. (2020) Drainage-structuring of ancestral variation and a common functional pathway shape limited genomic convergence in natural high- and low-predation guppies, *bioRxiv*.
Willing, E.-M., Bentzen, P., van Oosterhout, C., Hoffmann, M., Cable, J., Breden, F., Weigel, D. and Dreyer, C. (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies, *Molecular Ecology*, 19, pp. 968–984. doi: 10.1111/j.1365-294X.2010.04528.x.
Winge, Ø. (1937) Succession of broods in Lebistes, *Nature*, 140, p. 467.
Wittkopp, P. J., Williams, B. L., Selegue, J. E. and Carroll, S. B. (2003) Drosophila pigmentation evolution: Divergent genotypes underlying convergent phenotypes, *Proceedings of the National Academy of Sciences of the United States of America*, 100(4), pp. 1808–1813. doi: 10.1073/pnas.0336368100.
Wood, T. E., Burke, J. M. and Rieseberg, L. H. (2005) Parallel genotypic adaptation: when evolution repeats itself, *Genetica*, 123(1–2), pp. 157–170. doi: 10.1007/s10709-003-2738-9.
Wourms, J. P. (1981) Viviparity: The maternal-fetal relationship in fishes, *Integrative and Comparative Biology*, 21, pp. 473–515. doi: 10.1093/icb/21.2.473.
Wu, X. S., Masedunskas, A., Weigert, R., Copeland, N. G., Jenkins, N. A. and Hammer, J. A. (2012) Melanoregulin regulates a shedding mechanism that drives melanosome transfer from melanocytes to keratinocytes, *Proceedings of the National Academy of Sciences of the United States of America*, 109(31). doi: 10.1073/pnas.1209397109.
Zenger, K. R., Richardson, B. J. and Vachot-Griffin, A. M. (2003) A rapid population expansion retains genetic diversity within European rabbits in Australia, *Molecular Ecology*, 12(3), pp. 789–794. doi: 10.1046/j.1365-294X.2003.01759.x.
Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. and Yang, T.-L. (2018) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files, *Bioinformatics*, 35(10), pp. 1786–1788. doi: 10.1093/bioinformatics/bty875.
Zhou, Y. and Gaut, B. S. (2020) Large chromosomal variants drive adaptation in sunflowers, *Nature Plants*, 6(7), pp. 734–735. doi: 10.1038/s41477-020-0705-4.
Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D. and Gaut, B. S. (2019) The population genetics of structural variants in grapevine domestication, *Nature Plants*, 5(9), pp. 965–979. doi: 10.1038/s41477-019-0507-8.

SUPPLEMENTARY MATERIAL

Table S2.1 Fastsimcoal2 GHP/GLP parameter ranges used in the simulations of model 1 with migration.

Parameter	Uniform / loguniform	Lower bound	Upper bound	Fixed upper bound?
GHP Ne	logunif	1	50000	No
GLP ancestral NE	logunif	1	50000	No
GLP NE	logunif	1	50000	No
Time since divergence	unif	1	6E+07	No
GLP time to grow	unif	1	6E+07	No
GLP>GHP migration	logunif	1.00E-08	1	No
GHP>GLP migration	logunif	1.00E-08	1	No

Table S2.2 Fastsimcoal2 ECHP/ECLP parameter ranges used in the simulations of model 5 with migration.

Parameter	Uniform / loguniform	Lower bound	Upper bound	Fixed upper bound?
ECHP Ne	loguniform	1	50000	No
ECLP Ne	loguniform	1	50000	No
ECHP ancestral Ne	loguniform	1	50000	No
ECLP ancestral Ne	loguniform	1	50000	No
ECHP BN Ne	loguniform	1	50000	No
ECLP BN NE	loguniform	1	50000	No
Time since divergence	uniform	1	120	Yes
ECHP ancestral time	uniform	1	6.00E+07	No
ECLP>ECHP migration	logunif	1.00E-08	1	No
ECHP>ECLP migration	logunif	1.00E-08	1	No
ECLP BN start	Uniform	1	120	Yes
ECLP BN duration	Uniform	1	120	Yes
ECHP BN start	Uniform	1	6.00E+07	No
ECHP BN duration	uniform	1	6.00E+07	No

Table S2.3 Fastsimcoal2 GHP/TULP parameter ranges used in the simulations of model 1 without migration

Parameter	uniform / loguniform	Lower bound	Upper bound	Fixed upper bound?
TULP ancestral NE	loguniform	1	50000	No
TULP NE	loguniform	1	50000	No
GHP Ne	loguniform	1	50000	No
Time since divergence	Uniform	1	200	Yes
TULP time to grow	uniform	1	200	Yes

Table S2.4 Pairwise global F_{ST} values among all six populations with outlier windows removed from the data set. Mean F_{ST} below the diagonal, median F_{ST} above the diagonal. Population names can be found in table 2.1.

	GHP	GLP	ECHP	ECLP	TUHP	TULP
GHP		0.279	0.085	0.192	0.034	0.053
GLP	0.282		0.449	0.608	0.286	0.323
ECHP	0.104	0.435		0.056	0.114	0.144
ECLP	0.211	0.576	0.078		0.207	0.246
TUHP	0.040	0.291	0.127	0.220		0.030
TULP	0.063	0.328	0.158	0.259	0.036	

Table S2.5 Pairwise global F_{ST} values among all six populations with the sex chromosome removed from the data set. Mean F_{ST} below the diagonal, median F_{ST} above the diagonal. Population names can be found in table 2.1.

	GHP	GLP	ECHP	ECLP	TUHP	TULP
GHP		0.292	0.092	0.201	0.036	0.056
GLP	0.302		0.468	0.631	0.296	0.336
ECHP	0.120	0.458		0.060	0.120	0.152
ECLP	0.230	0.596	0.091		0.214	0.257
TUHP	0.046	0.311	0.141	0.237		0.032
TULP	0.073	0.350	0.176	0.279	0.041	

Table S3.1 Positions of the 41 outlier windows that overlap among three of the four measures in GHP/TULP.

Chromosome	BP1	BP2	# windows	Genes
chr1	10275001	10350000	1	MXI1, ADD3
	11625001	11775000	2	TEKT4, R3HCC1L, LOXL4, CRTAC1
	14700001	14925000	3	ALYREF, ARHGDIA, PPP1R12C, QPRT, ITGAL, ANAPC11, ZNF318, NKTR, PLA2G12B, PPP1R27, P4HB, PECAM1, RPTOR
	18675001	18750000	1	SMU1, DNAJA1, APTX, TRNAI-AAU, DSPP, SPARCL1, SSN6
	21375001	21450000	2	ANK3, CD027, PDE5A,
	21450001	21525000		PPP2R2B, WFS1, CRMP1, PCDH7
	27450001	27525000	1	PTP4A1, COL9A1, HYLS1, MRPLA, AHNAK, ABCC10, SP2
chr2	12675001	12750000	1	None
chr4	1800001	1875000	1	TRMT1L, PRPF38A, NIBAN, IFIT1, ZNF436
chr6	18075001	18150000	1	UNC45A, RCCD1, CIB1, GDPGP1, RHCG, SLS
chr7	18375001	18450000	1	B4GALT3, ATP6AP1, CACNA2D3, LRTM1
chr10	4575001	4650000	1	IL1RAPL1, B3GALT2
	6975001	7050000	1	None
chr12	1800001	1875000	1	CRB2, FBXW2, WDR45, NAS-4, ZCCHC9, XRCC4, VCAN
	10875001	11025000	2	PPEF2, AFF1, KLHL8, GDF7, MRC1, SDAD1, PTPN13, MAPK10,
chr15	4950001	5025000	1	CDH1, B-cadherin
chr16	34200001	34275000	1	TSNARE1
chr17	19350001	19425000	1	Uncharacterised, CCL25, RANBP3
chr18	1425001	1575000	2	RCE1, ACOX3, P2RY3, ALOX12, CD248, ZNF121, NAT16, MTM1, MTMR2
	12150001	12225000	1	None

	16575001	16650000	1	SEC24D, USP53, MYOZ2, SYNPO2
	18900001	18975000	1	PPP1R14B, PLCB3
chr20	675001	750000	1	C1QL3, DTX3L
	750001	825000	1	RSU1, TRDMT1
	1950001	2025000	1	CCR-1
000094F_0	525001	600000	1	PALMD, PLPPR4
	750001	1050000	4	SNX7, LOC103482394, LOC108165682
	1125001	1200000	1	None
	1275001	1350000	1	None
	1575001	1650000	2	None
000112F_0	225001	300000	1	NDUFC1, HARS, NAA15, MLS
000113F_0	675001	750000	1	PPFIA1

Table S3.2 Positions of the 118 outlier windows that overlap among three of the four measures in ECHP/ECLP.

Chromosome	BP1	BP2	# windows	Genes
chr1	31650001	31725000	1	None
chr2	24300001	24375000	1	KLF12
	24600001	24675000	1	MYCBP2, SCEL
chr4	23175001	23250000	1	GZMB
	26175001	26250000	1	PDGFRA, GSX2
	28425001	28575000	2	TPRG1L, LPP
	28950001	29025000	1	SLCO2A1, RYK, RAB6B
	29475001	29625000	2	SOX14, H2ST1
chr5	2325001	2400000	1	GRIP2, GGH
	6225001	6300000	1	RAP1GAP
	6375001	6525000	2	MAGI3
	11625001	11700000	1	HRH1, ATG7
chr7	5550001	5625000	1	None
	23250001	23325000	1	ZMYND8, TRNAK-CUU, TRAM1, LOC103467608
	25125001	25200000	1	CAMTA1
	26100001	26250000	2	CACNA1D, SYP, PLP2, PRICKLE3, FAM110A, MFSD2A, FKBP1A, SNPH, RAD21L1, OPN5
	26325001	26400000	1	VRP1, ATP1B1, YTHDF1, SLC17A9, GID8, DNAJC5
	26550001	26625000	1	OSBPL2, ADRM1, CABLES2, XKR7, RAP1A, KDM5BB

	27000001	27075000	1	PAN2, ORMDL2, NEMP1, CD63, LETMD1, GTSF1, HR38
chr8	6750001	6825000	1	LOC103468416, EMP1, GSG1, TCF20
	8850001	8925000	1	DESI1, FLII, LLGL1
	9300001	9375000	1	PRKCB
	10500001	10725000	3	NPTX1, PECAM1, PPP1R27, MCRIP1, ARHGDIA, RECQL5, ALYREF, CDR2, BTN3A3, LOC103469121
	11100001	11175000	1	EPS8, NME4, XPO6, FAM173A, HIRIP3, KCTD13, SEZ6L2
	12075001	12150000	1	CACNA1I
	12225001	12300000	1	TOB1
	12300001	12375000	1	SPAG9, LITAF
	12975001	13050000	1	HN1, SUMO2, NUP85, MRPS7, MIF4GD, SLC25A19, H2A2, RT1B
	13500001	13575000	1	FAM20C
	13650001	13725000	1	TBL3, RNF151, NDUFB10, RPL3L, TEX2
	13875001	13950000	1	TBC1D16, GAA, SAMD9L
	15150001	15225000	1	TRNAT-AGU, DNAJA3, HMOX2, LOC103460603, CCNF
	15375001	15600000	3	ADAP1, RAB26, SOX8, MED9, RASD1, USP22, SSTR2
	16575001	16650000	1	BAIAP2, CHMP6
	18000001	18075000	1	NOTCH2, RDH8
	18525001	18600000	1	DOCK7, LOC103468720, KANK2, HBA1, HBB, HBB-B1, HBZ,
	20550001	20625000	1	STAT5B, PLCL2
	21075001	21450000	5	NEUROD2, PPP1R1B, STARD3, MREG, TCAP, RPRML, ITGB3, METTL2A, TLK2, MRC2, RND2, VAT1, LOC103469365, PHB, ZNF652, EFTUD2, PHOSPHO1, GJC1
	25500001	25575000	1	RBFOX3
chr10	1050001	1125000	1	CSNK1A1, FBXO38
	3300001	3375000	1	None

	3525001	3600000	1	None
	5625001	5700000	1	NRG2
	9000001	9075000	1	ABLIM3, SH3TC2
	9150001	9225000	1	EBF1
	9750001	9900000	2	CTNNA1, LOC103471148
	10875001	11025000	2	FLT4, ERGIC1, DUSP1, NEURL1B, GPR50, VMA21
	12075001	12150000	1	GDF9, SOWAHA, SEPT8, CCNG1, NUDCD2, GABRA6, FGF1, NDFIP1, GNPDA1, HSPA4, UQCRQ
	13800001	13875000	1	TMCO6, ALDOB, HEPACAM, CNTN4, ARSI, XNP, M4A4D, RNF20
	19650001	19725000	1	RAB28
	20175001	20250000	1	GFRA4, SLC4A11
	24900001	24975000	1	GLRA1
	25200001	25275000	1	None
	26400001	26475000	1	FAT4
	26775001	26850000	1	SPRY1, SPATA5
	26850001	26925000	1	None
chr11	11625001	11700000	1	None
	22050001	22125000	1	PI4KB, FAM63A, PSBP1, LOC103472539, CTXN3
chr12	8775001	8925000	2	LAMC3
	9000001	9450000	6	LAMC3, AIF1L, PRKAA1, TTCC33, PTGER4, MAP1B, PTCD2, CACFD1, NOS1AP, GARNL3, PTGES, USP20, RPL7A, SURF1, SURF6, KYAT, CLG12HGORF114, ST6GALNAC6, ST6GALNAC4, TRNAI-AAU, FAM73B
	9825001	9900000	1	COL27A1B
	9975001	10200000	3	SEC16A, RPL28, PDCL, RPL37, ST8SIA5, LOXHD1, KCNT1, CAMSAP1, NOTCH1, PABPC3, PTGS1, RFP165
	10500001	10575000	1	ITGA1, KCMF1, NTRK2

	10875001	10950000	1	PPEF2, AFF1, KLHL8, SDAD1, GDF7, MRC1
	15600001	15750000	2	RNF34, LRRC75B, GGT1, GGT5, CABP1, GPAT4, RAB11FIP1
	16200001	16275000	1	KCNN2, SLC2A9, NT5C2
	18825001	18900000	1	NA
	19500001	19575000	1	PLPPR1
chr13	19275001	19350000	1	POLR2I, DACT3, PGRP-SC2, C5AR1
	27000001	27075000	1	LHFP
chr14	13350001	13425000	1	LOC103475867, TMEM88, PCOLCE2, MPP1, TMLHE
	22575001	22650000	1	GALNT10
chr15	7875001	7950000	1	None
	14175001	14250000	1	KCNMA1, DLG5
	15900001	15975000	1	LOXL4, CRTAC1, TTC31
	16200001	16275000	1	SLIT1, ABCC2
	16350001	16575000	3	CUTC, COX15, KCNIP2, HPS6, PROM1A, LDB1
chr17	1200001	1275000	1	NSUN4, UQCRH, DMBX1
	4950001	5025000	1	CLG17HXORF38
	5775001	5850000	1	RNF220
	25500001	25575000	1	PFKP, TRNAR-UCU, NCOA2
	27375001	27450000	1	ADCY8, EFR3A
chr18	2100001	2175000	1	None
	3825001	3900000	1	SORCS2, MRC2, VSIG1, LOC103480234, H1-1, H2B-I
chr20	9150001	9225000	1	DAP, ROPN1L, NUP58, MTMR6, ANKRD33B
	9375001	9525000	2	ZFHX4
chr23	1200001	1275000	1	CCT2, MYRFL, CTSE
	12825001	12900000	1	None
	13200001	13350000	2	GLIPR1, KRR1, PHLDA1, CMAH, LOC103459628, D2
000122F_0	525001	600000	1	MACROH2A1
000270F_0	75001	150000	1	PCDH10

Table S3.3 Positions of the 91 overlapping outlier windows among F_{ST}, |ΔAF| and D_{XY} in GHP/GLP

Chromosome	BP1	BP2	# windows	Genes
chr1	4725001	4800000	1	NA
	4950001	5025000	1	NA
	9000001	9075000	1	TUSC3
	19125001	19200000	1	NA
	19275001	19350000	1	NA
chr2	8175001	8250000	1	IL1RAPL1
	13050001	13125000	1	LOC108167039, KYNU, VWDE
	13725001	13800000	1	RNH1, NLRC3, DES, LOC103457762, GDF8, STAT1
	24900001	24975000	1	TFG, ABI3BP, IMPG2
	35475001	35550000	1	NA
	37725001	37800000	1	NA
	41175001	41250000	1	MYLK, OR10J4, UBE2G2, TSPEAR, TCHH
chr3	6075001	6150000	1	ADAMTS18, PHKB
	7425001	7500000	1	NA
	9000001	9075000	1	CLG3H15ORF59
	28050001	28125000	1	FTO
	28200001	28275000	1	IRX3
	30525001	30600000	1	LINGO1
chr4	17025001	17175000	2	CPN2, LRRC15, GP5, LOC103464132, RBM25, CLDN19, P3H2, CCDC181
	18900001	18975000	1	ELAVL4
	21450001	21525000	1	COLGALT2, CUNH1orf21, EDEM3
chr5	7650001	7725000	1	ARHGEF10L
	19725001	19800000	1	ZG57, LIMA1, SPRYD3, IGFBP3
	20325001	20400000	1	NAB2, DNAJC14, INPP1, S100B, ACOT9, SYS1, NRN1L, SSR1
	31725001	31800000	1	CNTN4
chr6	3225001	3300000	1	AASS, PTPRZ1
	12675001	12750000	1	SPIRE2, CHST6, TMEM231, GABARAPL2, ADAT1
	13050001	13125000	1	LOC103466360, OLFM4, NDRG4,

	17475001	17550000	1	NA
chr7	12975001	13050000	1	CNTN4
	25575001	25650000	1	RAP1GAP, NT5DC2
chr8	1350001	1425000	1	SEPT12
	7125001	7200000	1	CCDC134, MEI1
chr9	4800001	4875000	1	ARVCF
	12300001	12375000	1	NA
	13725001	13800000	1	XRCC4, TMEM167A, ARRDC3
	17475001	17550000	1	OPN4, putative defense protein Hdd11-like, EGFLAM, GDNF, NDC80
chr10	1125001	1200000	1	FBXO38, TMEM129
	1275001	1425000	2	RNF103, MAGT1
	2025001	2100000	1	NA
	17625001	17700000	1	MACROD1
chr11	1650001	1950000	4	THRB, NRD12, RPL15, NKIRAS1, UBE2E2, UBE2E1, SF3A3, MANEAL, PDIK1L, RPA2, FOXO6
	2700001	2775000	1	LOC103471987
chr12	3000001	3075000	1	AUH, SYK
	3150001	3225000	1	CHD1
chr13	6600001	6675000	1	RTN4RL1
	17850001	17925000	1	NA
	24075001	24150000	1	EIF4G1, PSMD2, K12H4.7, SERPINE2
chr14	20400001	20475000	1	ZBTB16, HTR3A, HTR3B, USP28, ORAI2
	22050001	22125000	1	KLF5
	22800001	22950000	2	GALNT10
chr15	4950001	5025000	1	CDH1, B-CADHERIN
	8100001	8175000	1	MRPS5, TRNAI-UAU, TCHH1L, MAL, NPHP1
	12225001	12300000	1	FAM196A
	12675001	12750000	1	SORCS1
	16725001	16800000	1	ANTXR1, PRDM8, GDF10, PTPN13
	20325001	20400000	1	LOC103477205, NSMCE4A, CCDC177, SLC39A9
chr16	10050001	10125000	1	MAN1A1
	17250001	17325000	1	SEC61B, ALG2, FRRS1L, TMEM245

	22425001	22500000	1	NA
chr17	11100001	11250000		CAPN10, EIF4E2, PLSCR5, CHRNG, CCL20, PRSS56, VWA5B2, ALG3, MUL1, METTL13
	11775001	11850000	1	EPHB1
	14250001	14325000	1	CHD2
chr18	15000001	15150000	2	ELAVL2
chr20	1	75000	1	ARHGEF4
	300001	525000	3	STEAP2, MOS, ALG14, CNN3, F3, ABCD3, CUBN
	600001	750000	2	CUBN, DTX3L
	2625001	2775000	2	PIM1
	4125001	4200000	1	NRP1
	16575001	16650000	1	TYMS, ENOSF1, YES1, ADCYAP1
chr22	13425001	13500000	1	TSSC1, TRAPPC12, ADI1
	14400001	14475000	1	COL12A1, TMEM30A, FILIP1, CLDN20
chr23	10125001	10200000	1	DOCK4
	12150001	12225000	1	CELSR1, TRMU, TDCB, TPRKB, ALG10, LOC103459547
000117F_0	675001	750000	1	AKAP6
	825001	1125000	4	AKAP6, NPAS3

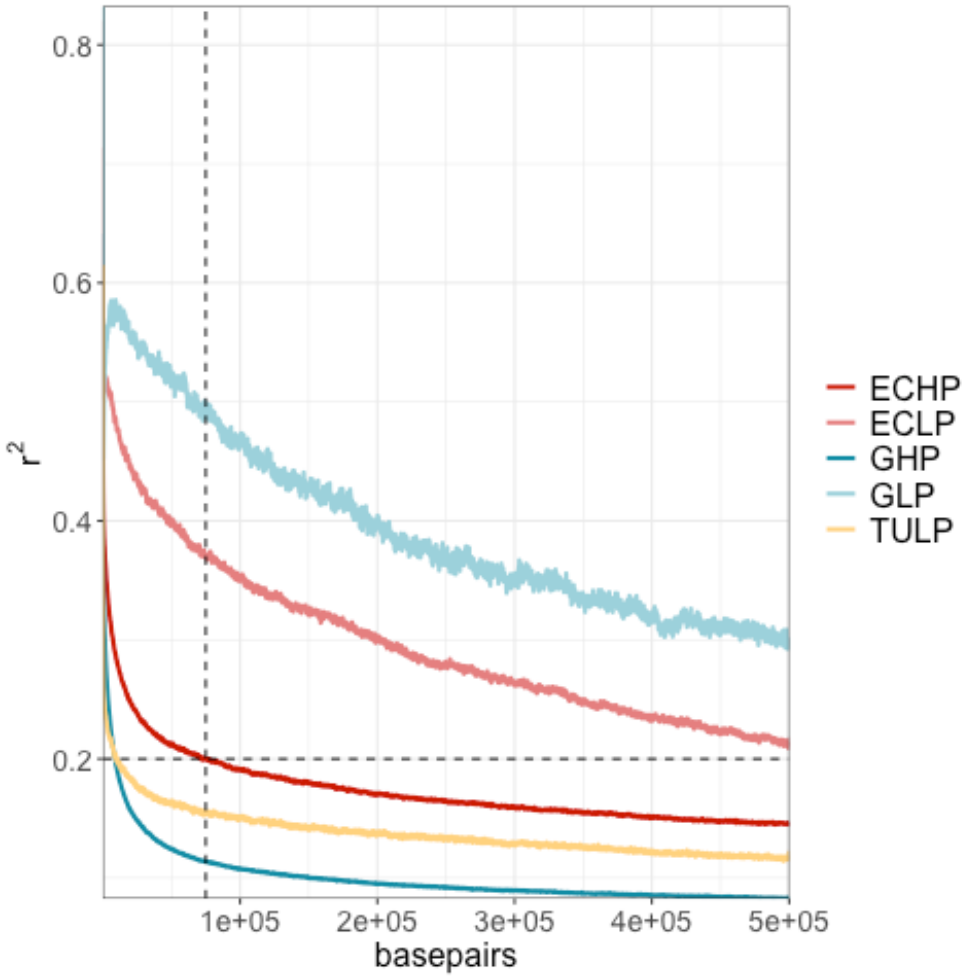


Figure S3.1 LD decay analysis to determine window size for the genome scan analyses. Vertical dashed line indicates the selected window size, horizontal dashed line the cutoff for LD decay.

Table S4.1 Fastsimcoal2 parameter ranges used in the simulations for all four populations.

Parameter	Uniform / loguniform	Lower bound	Upper bound
HP NE	loguniform	1	50000
LP NE	loguniform	1	50000
LP ancestral NE	loguniform	1	50000
LP BN Ne	loguniform	1	50000
HP/LP divergence	uniform	1	30
LP>HP migration	loguniform	1.00E-08	1.00E-02
HP>LP migration	loguniform	1.00E-08	1.00E-02
LP BN start	uniform	1.00	30.00
LP BN duration	uniform	1	30

Table S4.2 Pairwise global F_{ST} values among all six populations with outlier windows removed from the data set. Mean F_{ST} below the diagonal, median F_{ST} above the diagonal. Population names can be found in table 4.1.

	ILL	IUL	IC	IT	GLP	GHP
ILL		0.015	0.022	0.028	0.257	0.011
IUL	0.022		0.021	0.029	0.263	0.011
IC	0.030	0.026		0.025	0.279	0.013
IT	0.036	0.036	0.033		0.270	0.021
GLP	0.264	0.269	0.287	0.278		0.299
GHP	0.019	0.017	0.019	0.028	0.299	

Table S4.3 Pairwise global F_{ST} values among all six populations with the sex chromosome removed from the data set. Mean F_{ST} below the diagonal, median F_{ST} above the diagonal. Population names can be found in table 4.1.

	ILL	IUL	IC	IT	GLP	GHP
		0.017	0.024	0.031	0.269	0.013
ILL	0.028		0.022	0.032	0.271	0.013
IUL	0.036	0.032		0.027	0.290	0.014
IC	0.044	0.043	0.040		0.281	0.022
IT	0.284	0.286	0.305	0.298		0.309
GLP	0.025	0.022	0.024	0.035	0.317	

Table S4.4 FASTSIMCOAL2 parameter estimates from the best run and 95% confidence intervals generated by simulations of the best model in ILL. Population size estimates are given in diploid number of animals, and times are given in generations. Migration rates are the probability of a gene to move from one population to other per generation. Growth rates are given per generation, positive values indicate population contraction forward in time. BN = bottleneck. Values marked with * indicate estimates that do not fit within the simulated 95% confidence interval. (table on the next page).

A. ILL/GHP model 2M	Point estimate	Lower bound	Upper bound
HP N_E	*138	375	3650
LP N_E	*47	11	27
LP ancestral N_E	31205	27880	34847
LP BN LP	*501	1155	4992
HP/LP divergence	7	7	7
LP growth	*0.558	1.174	1.609
LP>HP migration	*2.27E-04	2.53E-05	2.13E-04
HP>LP migration	1.48E-06	4.95E-08	2.15E-06
LP BN start	7	7	7
LP BN end	2	2	2

B. IUL/GHP model 4M	Point estimate	Lower bound	Upper bound
HP N_E	545	418	769
LP N_E	31	22	50
LP ancestral N_E	22757	12003	25304
LP BN N_E	646	389	1147
HP/LP divergence	27	27	27
LP growth	1.52	1.16	1.93
HP growth	-1.16E-07	-3.71E-07	1.07E-07
LP>HP migration	2.27E-04	3.71E-05	7.49E-05
HP>LP migration	1.48E-06	6.28E-08	1.45E-05
LP BN start	27	27	27
LP BN end	1	1	1

C. IC/GHP model 4M	Point estimate	Lower bound	Upper bound
HP N_E	9652	8744	9652
LP N_E	637	357	682
LP ancestral N_E	13	6	24
LP BN N_E	788	522	871
HP/LP divergence	71	61	89
LP growth	1.33E-02	-9.89E-03	4.32E-02
HP growth	-9.49E-08	-6.41E-07	-2.75E-08
LP>HP migration	1.60E-02	1.07E-02	1.96E-02
HP>LP migration	1.97E-02	1.29E-02	2.64E-02
LP BN start	44	36	54
LP BN end	15	10	32

D. IT/GHP model 4M	Point estimate	Lower bound	Upper bound
HP N_E	11608	10498	11608
LP N_E	9	7	48
LP ancestral N_E	2	1	3
LP BN NE	1189	376	1892
HP/LP divergence	86	39	90
LP growth	1.65	0.14	1.74
HP growth	-1.66E-07	-6.49E-07	-3.40E-08
LP>HP migration	1.33E-02	9.71E-03	2.40E-02
HP>LP migration	2.37E-02	1.59E-02	3.17E-02
LP BN start	23	22	35
LP BN end	2	2	15

Supplementary table S5.1 XP-EHH outlier windows for ILL

Chromosome	BP1	BP2	number of windows
chr1	13200001	13275000	1
chr3	23175001	23250000	1
chr4	3525001	3750000	3
chr5	20700001	20775000	1
	22275001	22350000	1
	25950001	26025000	1
	26175001	26250000	1
chr6	27300001	27375000	1
	27525001	27600000	1
chr9	2175001	2250000	1
	2400001	2475000	1
	18075001	18150000	1
	18525001	18600000	1
	19950001	20100000	2
	21075001	21150000	1
	21225001	21600000	5
	23325001	23400000	1
	23475001	23550000	1
chr11	675001	750000	1
	7725001	7800000	1
chr14	3750001	3825000	1
chr15	1875001	1950000	1
	4950001	5100000	2
	5400001	5550000	2
	6225001	6300000	1
	6600001	6675000	1
chr16	6375001	6450000	1
chr16	13425001	13500000	1
chr18	7275001	7350000	1

Supplementary table S5.2 XP-EHH outlier windows for IUL

Chromosome	BP1	BP2	number of windows
chr2	13950001	14025000	1
chr5	22875001	22950000	1
	26100001	26250000	2
chr16	8925001	9000000	1
	23175001	23250000	1
	23775001	23850000	1
	25050001	25125000	1
chr18	21825001	21900000	1
	26250001	26325000	1
chr19	20475001	20700000	3
	20775001	20925000	2
	21150001	21225000	1
chr21	17625001	17700000	1

Supplementary table S5.3 XP-EHH outlier windows for IC

Chromosome	BP1	BP2	number of windows
chr1	7650001	7725000	1
chr2	225001	300000	1
	30375001	30450000	1
	41025001	41100000	1
chr5	26100001	26175000	1
chr11	24750001	24825000	1
chr12	1200001	1275000	1
chr13	5700001	5850000	2
chr14	5625001	5700000	1
chr15	5025001	5550000	5
	5775001	5850000	1
	6075001	6150000	1
	6225001	6300000	1
	26400001	26475000	1
chr20	1500001	1575000	1

Supplementary table S5.4 XP-EHH outlier windows for IT

Chromosome	BP1	BP2	number of windows
chr1	7275001	7350000	1
	27900001	27975000	1
	28575001	28650000	1
chr3	1125001	1200000	1
chr4	3600001	3900000	4
	3975001	4125000	2
	4200001	4275000	1
	9300001	9375000	1
chr5	16950001	17025000	1
	17475001	17550000	1
	25575001	25725000	2
	28050001	28125000	1
chr8	21375001	21450000	1
	22575001	22650000	1
chr9	31800001	31875000	1
chr11	4500001	4575000	1
chr13	25500001	25575000	1
chr14	3075001	3150000	1
	7725001	8100000	5
	8325001	8400000	1
chr15	4125001	4800000	9
	5025001	5100000	1
	5325001	5625000	4
	5700001	5775000	1
	6600001	6675000	1
	26325001	26475000	2
chr17	19650001	19800000	2
	24675001	24750000	1
	26625001	26700000	1
chr18	21825001	21900000	1
	21975001	22050000	1
chr21	19050001	19200000	2
chr23	6075001	6150000	1
000211F_0	150001	225000	1

Supplementary table S5.5 iHH12 outlier windows for ILL

Chromosome	BP1	BP2	number of windows
chr1	14250001	14325000	1
chr3	21750001	21825000	1
chr4	3450001	3525000	1
chr5	22200001	22275000	1
	26025001	26100000	1
chr9	20625001	20700000	1
	21150001	21525000	5
	23025001	23100000	1
	23325001	23475000	2
	27750001	27825000	1
chr15	5175001	5550000	5
chr18	11325001	11475000	2
000104F_0	1200001	1275000	1

Supplementary table S5.6 iHH12 outlier windows for IUL

Chromosome	BP1	BP2	number of windows
chr1	10875001	10950000	1
	13725001	13875000	2
	14250001	14325000	1
chr2	40650001	40725000	1
chr5	26025001	26100000	1
chr8	6450001	6525000	1
chr9	27750001	27825000	1
chr11	19050001	19125000	1
chr15	21600001	21675000	1
chr15	24975001	25050000	1
chr19	20475001	20625000	2
	20775001	20850000	1
chr23	5775001	5850000	1

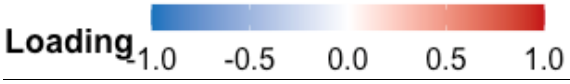
Supplementary table S5.7 iHH12 outlier windows for IC

Chromosome	BP1	BP2	number of windows
chr1	11925001	12000000	1
	13800001	13875000	1
chr2	2325001	2400000	1
chr3	19725001	19800000	1
chr4	27525001	27675000	2
chr6	9450001	9525000	1
	16500001	16575000	1
chr9	27750001	27825000	1
chr10	24375001	24450000	1
chr13	5700001	5775000	1
	27750001	27825000	1
chr14	4950001	5025000	1
chr15	5175001	5475000	4
chr16	10950001	11025000	1
chr17	18225001	18300000	1
chr19	18000001	18075000	1
chr22	4575001	4650000	1
	14850001	14925000	1
chr23	1575001	1650000	1
000104F_0	1200001	1275000	1
000184F_0	150001	225000	1

Supplementary table S5.8 iHH12 outlier windows for IT

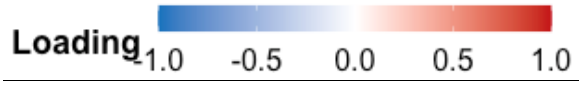
Chromosome	BP1	BP2	Number of windows
chr1	13725001	13800000	1
	27750001	28200000	6
chr4	3450001	3675000	3
	3975001	4050000	1
chr5	12300001	12375000	1
	24900001	24975000	1
	31200001	31275000	1
chr6	16500001	16575000	1
chr9	20625001	20700000	1
chr10	24375001	24525000	2
	26175001	26250000	1
chr11	4425001	4500000	1
chr13	22275001	22350000	1
	27750001	27825000	1
chr14	5250001	5400000	2
	7800001	7875000	1
	7950001	8025000	1
	8850001	8925000	1
	10050001	10125000	1
chr15	5100001	5550000	6
	5625001	5700000	1
	26325001	26400000	1
chr16	16425001	16500000	1
chr17	18225001	18300000	1
chr21	19050001	19125000	1
chr23	6600001	6675000	1

Supplementary table S5.9 Windows with outlier eigenvalues on eigenvector 1.



Chrom	BP1	BP2	Eigen-value	Loadings on eigenvector 1 ILL	IUL	IC	IT
chr3	33221787	33313394	3.589	-0.502	-0.508	-0.483	-0.506
	33313541	33353061	3.734	-0.500	-0.506	-0.506	-0.488
chr4	6402404	6436726	3.683	-0.499	-0.495	-0.499	-0.507
chr5	20678404	20744930	3.726	-0.501	-0.502	-0.493	-0.503
chr8	9570352	9611892	3.668	-0.493	-0.498	-0.503	-0.507
	24470692	24512621	3.564	-0.503	-0.499	-0.490	-0.508
	24584590	24629724	3.615	-0.502	-0.503	-0.484	-0.511
	24629728	24674402	3.560	-0.513	-0.490	-0.501	-0.496
	24690959	24707669	3.593	-0.497	-0.496	-0.498	-0.508
	24785037	24817048	3.889	-0.502	-0.495	-0.502	-0.501
	24897445	24926067	3.580	-0.505	-0.500	-0.514	-0.481
	25036898	25073937	3.798	0.497	0.499	0.503	0.501
chr10	917992	1044113	3.622	-0.503	-0.511	-0.484	-0.502
	1044123	1141168	3.615	-0.498	-0.512	-0.487	-0.502
	1235773	1336327	3.671	-0.498	-0.509	-0.490	-0.503
	1336357	1383077	3.726	-0.494	-0.506	-0.495	-0.504
	1383210	1439400	3.641	-0.501	-0.508	-0.488	-0.502
	1439475	1579703	3.583	-0.505	-0.512	-0.487	-0.496
	1579704	1686506	3.615	-0.506	-0.513	-0.483	-0.497
	1686564	1777672	3.647	-0.504	-0.512	-0.482	-0.501
	1777799	1857611	3.616	-0.507	-0.515	-0.479	-0.497
chr13	5286314	5344456	3.733	0.505	0.503	0.498	0.494
	5344571	5401361	3.642	-0.503	-0.509	-0.476	-0.511
	9270799	9353146	3.755	-0.499	-0.507	-0.505	-0.489
chr14	3062057	3104901	3.692	-0.504	-0.491	-0.504	-0.500
chr15	4938877	4985391	3.718	-0.504	-0.480	-0.505	-0.510
	5021202	5060367	3.811	-0.504	-0.489	-0.502	-0.505
	5060423	5099719	3.933	0.501	0.496	0.502	0.502
	5222773	5281854	3.805	-0.504	-0.483	-0.507	-0.505
	5350885	5412354	3.742	-0.497	-0.488	-0.507	-0.508
	5412384	5462574	3.780	-0.493	-0.494	-0.505	-0.508
	6587917	6651087	3.640	-0.496	-0.502	-0.499	-0.504
	24785567	24890042	3.678	-0.499	-0.489	-0.510	-0.502
chr21	1907512	1949195	3.685	-0.484	-0.505	-0.502	-0.509
chr23	3791263	3861636	3.526	-0.484	-0.500	-0.507	-0.508
	6848531	6908411	3.476	-0.489	-0.481	-0.504	-0.525
	9411798	9494372	3.513	-0.466	-0.503	-0.512	-0.517
	11675442	11739813	3.562	-0.502	-0.502	-0.500	-0.495

Supplementary table S5.10 Windows with outlier eigenvalues on eigenvector 2.



				Loadings on eigenvector 1			
Chrom	BP1	BP2	Eigen-value	ILL	IUL	IC	IT
chr1	25099485	25133126	1.598	0.583	0.683	-0.319	-0.303
	32315406	32364164	1.577	-0.635	-0.696	-0.239	-0.234
chr2	11175085	11229877	1.696	-0.205	-0.339	0.702	0.592
chr3	4230794	4333325	1.687	-0.702	0.196	-0.565	0.387
	14472356	14530026	1.664	-0.261	-0.609	0.336	0.670
chr7	5395097	5430515	1.676	0.091	-0.205	0.668	0.710
chr12	21784928	21817261	1.621	0.266	0.353	-0.662	-0.605
	21844016	21870625	1.671	-0.544	-0.484	-0.539	-0.424
chr13	17604662	17642008	1.689	0.235	0.353	-0.555	-0.715
chr14	12237288	12290352	1.623	0.671	0.701	-0.012	-0.242
	12405134	12459704	1.607	-0.720	-0.591	-0.267	-0.249
	12459788	12505696	1.583	-0.465	-0.573	0.398	0.545
	19741307	19796628	1.662	-0.659	-0.676	-0.107	-0.313
	21297239	21360465	1.612	-0.693	-0.707	0.134	-0.043
chr17	9289418	9328144	1.642	0.485	0.493	-0.489	-0.532
	21954450	22031525	1.609	0.473	0.695	0.537	0.070
chr18	26292405	26312060	1.681	0.324	-0.695	-0.619	0.168
chr19	21998888	22070753	1.559	-0.610	0.333	-0.568	0.441