UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA VEGETAL

Analysis of genomic changes during adaptive evolution in *Drosophila subobscura* populations of contrasting biogeographical history

Marta Maria Alves Antunes

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Dissertação orientada por:
Margarida Maria Demony de Carneiro Pacheco de Matos
Sofia Gonçalves Seabra

2017

Index

Agradecimentos

Com a conclusão desta etapa do meu percurso académico, não podia deixar de agradecer a todas as pessoas que me apoiaram e contribuíram para a realização deste trabalho que foi o culminar de dois anos intensos em termos de aprendizagem.

Em primeiro lugar gostava de agradecer às minhas orientadoras, a professora Margarida Matos e a professora Sofia Seabra, por toda a ajuda, paciência, apoio e disponibilidade. É com muita certeza que afirmo que poucas pessoas fariam o que elas fizeram, fizeram-me sentir parte da família, ajudaram-me em tantas coisas e de tantas formas que era impossível enumerá-las aqui. Queria também agradecer ao Pedro Simões por tudo o que me ajudou neste trabalho e à Inês Fragata que, embora já não esteja a trabalhar connosco, continua a fazer parte desta equipa porque está sempre pronta a discutir resultados e a dar e ouvir ideias acerca de novas experiências.

Um agradecimento especial queria dirigir ao Francisco-Pina Martins pela sua disponibilidade constante para me ajudar a lidar com o sistema operativo LINUX, à Ana Quina e à Ana Vieira pela ajuda na análise de Pool-sequencing e ao Mauro Santos pela sua intervenção na discussão sobre Linkage Disequilibrium. De não esquecer também o Vítor Sousa que tanto ajudou o nosso grupo.

Queria agradecer também a todos os professores que me deram aulas ao longo do ano curricular do mestrado pois foram eles que me iniciaram nesta viagem. Em particular gostaria de mencionar o professor Francisco Couto que me deu a oportunidade de colaborar em vários projectos de investigação onde trabalhei com a Márcia Barros e com o André Lamúrias. Um obrigado a eles também.

Em termos mais pessoais gostaria de agradecer à minha família e namorado Rúben Martins pelo apoio incondicional, principalmente nos momentos mais difíceis. Queria agradecer aos meus amigos, em especial às minhas queridas amigas da residência Laranjeiras, à Arsénia Massinga, Patrícia Abrantes, Telma Varela, Maria Martins, Sandrina Teixeira e Ana Rita Libório. Também queria agradecer à Dona Elvira e à equipa dos seviços informáticos da Faculdade de Ciências.

Finalmente gostaria de agradecer o financiamento deste projecto. As sequências analisadas nesta tese de mestrado foram obtidas em trabalhos anteriores apoiados pelos Fundos nacionais portugueses através da Fundação para a Ciência e Tecnologia, projeto PTDC / BIA-BIC / 2165/2012. Agradeço também ao Centro de Ecologia, Evolução e Alterações Ambientais (cE3c) e ao seu financiamento da FCT UID / BIA / 00329/2013 pelo suporte logístico e financeiro em comunicações extra-murais do trabalho realizado ao longo do ano.

Resumo

Um dos objectivos principais da biologia é entender como é que os organismos evoluem e se adaptam a novos ambientes. A prática de evolução experimental em laboratório é uma metodologia muito eficaz pois permite estudar a evolução a ocorrer em tempo real. Em alternativa a estudar a evolução na natureza, muitas vezes dificil de realizar, as populações são analisadas ao longo de gerações de evolução em ambiente controlado do laboratório. Desta forma, é muito mais fácil perceber quais as variáveis que podem estar envolvidas no processo evolutivo e evitar efeitos imprevisíveis exteriores à experiência. Também é possível manusear os indivíduos e controlar o seu acasalamento, garantindo desta forma a não ocorrência de troca de genes não pretendida entre populações. O aparecimento da possibilidade de sequenciar muitos indíviduos a baixos custos permitiu explorar esta temática da evolução de populações também ao nível genómico. Surgiram nomeadamente as técnicas de sequenciação genómicas de conjuntos ("pools") de indíviduos ("Pool-sequencing") e sequenciação de DNA associado a locais de restrição ("RAD-sequencing"). Estas duas técnicas foram usadas na obtenção de dados analisados neste trabalho. A abordagem usada chama-se "evolve and resequencing", ou seja "evoluir e resequenciar", e significa que as populações são sequenciadas em vários momentos ao longo de diversas gerações, de forma a ser possível acompanhar alterações evolutivas que estejam a ocorrer no genoma dos indivíduos.

Neste trabalho foram analisadas as alterações genómicas ocorridas em duas populações de *Drosophila subobscura* amostradas em dois locais europeus de latitudes contrastantes (Adraga, em Sintra, Portugal e Groningen na Holanda) durante a sua adaptação ao ambiente do laboratório. A escolha desta espécie deveu-se ao facto de apresentar elevado polimorfismo ao nível de inversões cromossómicas no seu genoma e já ter sido observada, para as duas populações referidas anteriormente, convergência a nível fenotípico quando colocadas em laboratório mas não a nível da frequência das inversões.

Os arranjos cromossómicos (inversões) foram um destaque neste trabalho porque afectam a arquitectura genómica das populações ao reduzirem a recombinação nos locais onde se encontram. Foram indicados como também estando implicados na adaptação climática nesta espécie dado que a sua frequência varia de acordo com a latitude onde se encontram as populações que as possuem.

Na primeira parte do trabalho, caracterizei os polimorfismos nucleotídicos simples (SNPs) obtidos no estudo de "Pool-sequencing" que deram indicações de seleção, analisando os que foram reconhecidos como estando associados (i.e. que deram "hits") a proteínas nestas duas populações. Detectei que muitos genes estão sob seleção nestas populações o que sugere uma base poligénica de adaptação. Também observei que estão envolvidos em processos biológicos distintos entre populações, reforçando a constatação de que as populações apesar de convergirem fenotipicamente o fazem por caminhos genéticos distintos. A única família de genes que foi encontrada sob seleção nas duas populações foi a famíla de genes do receptor gustativo, envolvido no reconhecimento de alimentos. Também caracterizei o tipo de mutação que cada SNP provoca e foi interessante descobrir que alguns dos *SNPs* sob seleção se encontram em pequenas regiões intrónicas.

A segunda parte do trabalho consistiu em analisar os dados obtidos por "RAD-sequencing". As duas metodologias de sequenciação são complementares uma vez que a sequenciação genómica de "pools" de indivíduos permite obter mais marcadores de DNA mas sem informação individuals enquanto que usando enzimas de restrição se obtêm menos marcadores mas com informação

individual para muitos indivíduos separadamente. Esta última abordagem, permitiu observar que os indivíduos sequenciados estão mais separados pelas inversões que possuem do que pela população a que pertencem, ou seja, um indíviduo de uma população pode ser mais semelhante a um de outra população se possuir a mesma inversão no seu genoma.

Além disso, foram observadas as alterações genómicas em cromosomas com inversões específicas, tanto ao nível do total de SNPs detectados em cada cromosoma como também ao nível dos SNPs sob seleção. Como ilustração das potencialidades deste estudo, neste trabalho foram analisadas três arranjos cromossómicos, O_{3+4}, O_{ST} e A_2. Foram analisadas alterações entre gerações da mesma população mas também a diferenciação entre populações e como esta diferenciação entre populações evolui ao longo do tempo.

As análises dos cromosomas com a inversão A2 foram aquelas que permitiram uma mais robusta análise dos resultados devido ao número de indivíduos amostrados. Foi detectada maior diferenciação entre gerações nos SNPs da população de Groningen do que nos de Adraga, o que está de acordo com o que foi encontrado na análise de dados de Pool-seq onde a maioria dos SNPs candidatos nesta população se encontram no cromosoma A. Isto poderá indicar que o cromosoma A tem um papel fundamental na adaptação desta população ao ambiente do laboratório. O total de SNPs neste cromossoma não deu indicações de convergência entre as duas populações, pelo contrário sugerem divergência, i.e. aumento da diferenciação entre as populações ao longo do tempo. Um aspecto fundamental foi analisar até que ponto as populações dão indicação de terem uma dinâmica adaptativa semelhante a nivel genómico, ou se, pelo contrário, elas não convergem em SNPs com sinal de selecção. De facto, em concordância com dados de pool-seq, na análise dos cromossomas com inversão A2 foram detectados poucos SNPs comuns entre as duas populações a darem sinal de selecção. Como era de esperar, a diferenciação dos SNPs candidatos (i.e. aqueles que deram sinal de selecção) entre gerações foi maior na população para a qual os SNPs foram detectados. No entanto, os mesmos SNPs também responderam na outra população com alterações temporais superiores à diferenciação global de todos os SNPs, o que sugere que os SNPs que estão sob seleção numa populacão também podem estar, pelo menos em parte, na outra.

Comparando os resultados obtidos nas várias inversões/arranjos analisados, podemos dizer que não foi detectada convergência nem para o total de SNPs em cada cromosoma nem entre os SNPs sob seleção em cada cromosoma, o que indica que, em geral, estas populações usam diferentes vias a nível genómico para atingir o mesmo estado a nível fenotípico.

Este trabalho vem adicionar novos elementos para a questão dos mecanismos que levam à manutenção das inversões. Permite perceber que esta é uma questão complexa e portanto é necessária a realização de mais análises, nomeadamente análises de diferenciação nos outros cromossomas, análises de desequilíbrio de ligação (*"*linkage disequilibrium*"*) e o mapeamento de mais zonas do genoma.

Finalmente, ao nível de ferramentas bioinformáticas criadas, foram desenvolvidas duas que foram essenciais para a realização deste trabalho. Especificamente desenvolvi: uma base de dados que denominei *DsubSeqLoc*, que integra informação que estava dispersa na literatura, facilitando a localização cromossómica de sequências de genes de *Drosophila subobscura* já publicados, relevante quer em análises de dados de "Pool-seq" quer nos dados de "RAD-seq"; e uma "pipeline" que permite remover um dos haplótipos parentais do genótipo dos descendentes, necessário para a análise genómica das sequencias de RAD-seq.

A base de dados *DsubSeqLoc* permite armazenar num único local toda a informação que se conhece atualmente acerca da localização de sequências de *Drosophila subobscura.* A integração desta informação é essencial pois ainda não existe um genoma de referência completamente anotado para a espécie e vai crescer em tamanho e importância à medida que novas sequências vão sendo mapeadas. Prevê-se que esta base de dados vá contribuir muito para futuros estudos nesta espécie. Além disso, também foi criada uma página *web* (http://www-personal.fc.ul.pt/~mmmatos/DsubSeqLoc) que permite um fácil acesso à informação mesmo a utilizadores que não estejam familiarizados com a escrita de *queries* SQL. Por sua vez a "pipeline" criada no contexto desta dissertação permite remover um dos haplótipos parentais dos genótipo dos seus descendentes. No contexto desta tese foi uma ferramenta muito util na análise de dados de RAD-seq, pois o DNA extraído pertencia a larvas que resultaram do cruzamento das nossas populações com a linha homocariotípica *chcu* (protocolo usado na identificação das inversões), e apenas queríamos, obviamente, analisar os haplótipos das nossas populações. Por exemplo, no contexto deste trabalho, o DNA extraído pertencia a larvas que resultaram do cruzamento de uma população *wild* com a linha homocaritípica *chcu*, mas apenas queriamos analisar os haplótipos das populações *wild.* A pipeline serviu deste modo para simplificar o processo de eliminação dos haplótipos que não tinham interesse para o nosso estudo. Se não tivesse sido criada muitos programas individuais teriam de ser executados para obter os haplótipos de interesse. Além disso, a pipeline permite fazer um passo de filtragem e mantém os registos da quantidade de dados que foi filtrada. Além de tudo isto, a pipeline não apresenta especificidade para os nossos dados, podendo ser utilizada na realização de outros estudos. Estará assim disponivel a outros utilizadores, e.g. sempre que seja requerido retirar o haplótipo de um dos progenitores.

Este trabalho permitiu entender melhor o processo de adaptação ao nível genómico de populações de *Drosophila subobscura.* Este trabalho pioneiro abriu novos horizontes de investigação, deixando interessantes questões em aberto a abordar no futuro.

Palavras-chave: *Drosophila subobscura*, evolução e resequenciação, base de dados para localização de sequências, inversões cromossómicas, "RAD-sequencing"

Abstract

Experimental evolution is a powerful approach to study adaptation of populations in real-time. Using this approach, I studied at the genome-wide level the evolution of two populations of *Drosophila subobscura* derived from two contrasting biogeographical latitudes (Adraga, Portugal and Groningen, Netherlands), across generations since laboratory introduction. Modern sequencing technologies are now providing a high resolution in the analysis of patterns of genetic variation. In the context of this dissertation, I analyzed 'evolve-and-resequence' data (that is, genomic information across generations) of both populations obtained by Pool-sequencing and RAD-sequencing at two generations (6 and 25) after founding these populations in the common, laboratorial environment.

With the pool-seq data I characterized SNPs that indicated selection and gave hits with proteins. I discovered that many genes and different biological processes are at play, suggesting a polygenic basis of adaptation. Only one family of genes was found in common between the two populations, associated with recognition of taste stimuli. I also classified each SNP in the type of mutation and interestingly found several genes under selection in small intronic regions.

Chromosomal inversions may play an important role in genomic evolution and population differentiation, because they affect the genomic architecture of populations by suppressing or reducing recombination in these inverted regions. In the context of this thesis I analysed the RAD-sequencing data of many individuals with known karyotypes of both populations across generations. Interestingly, individuals were more clearly separated by the inversion they carry than by the population to which they belong. I compared the differentiation between generations and between populations both for all SNPs in each chromosome and for candidate SNPs (that gave signs of being under selection). Also, I analysed how the genetic differentiation between populations changed through time. I detected that candidate SNPs differed between populations, in accordance with what was already observed at the Pool-seq level. Nevertheless, the SNPs under selection in one population also suggested some selective response in the other, although to a smaller extent. It was not detected convergence between the two populations neither for total SNPs of each chromosome neither for candidate SNPs. Focusing on the chromosomes with A2 inversion there was a higher differentiation between populations than chromosomes with other inversions, considering the same period of time.

Furthermore, I developed two bioinformatic tools that were essential to make the analyses: a database called *DsubSeqLoc* (http://www-personal.fc.ul.pt/~mmmatos/DsubSeqLoc) that integrates information already published of the cytological location of genes or of other genomic regions in *Drosophila subobscura* and a pipeline that removes one parental haplotype from the progeny.

Keywords: *Drosophila subobscura*, evolve and resequence, sequences location database, chromosomal inversions, RAD-sequencing

List of Figures

List of Tables

List of abbreviations

We adopted the following nomenclature:

Ad - refers to population sampled in Adraga (Portugal) in 2010
Gro - refers to population sampled in Groningen (Netherlands) in 2010
G6 – Generation 6
G25 - Generation 25
PCA – Principal component analysis
PCoA - Principal coordinates analysis
Pool-seq – Pool-sequencing
RAD-seq – RAD-sequencing
LD – Linkage Disequilibrium
FISH - Fluorescence In Situ Hybridization
SNPs - Single Nucleotide Polymorphisms
ER – Entity Relationship
F_{ST} - Fixation Index
FDR - False Discovery Rate

1. Introduction

1.1 Genomic approaches to study adaptive evolution in experimental evolution

Adaptive evolution - evolutionary changes that are adaptive - occurs not randomly but as a consequence of the changes in the genetic constitution of a population due to natural selection acting in specific features of the environment (Merilä and Hendry 2014). Such changes increase fitness of the individuals by addressing some specific challenges presented by the environment.

Experimental Evolution is a powerful approach to follow the previous mentioned changes in real time (Buckling et al. 2009). In particular the 'evolve and resequence' approach (Turner et al. 2011, Baldwin-Brown et al. 2014, Long et al. 2015, Schlötterer et al. 2015) revolutionized the studies in genomic evolution by analyzing the changes of DNA sequences in several time points in the course of an evolutionary process.

This is possible due to the increase in computational power as well as the arrival of several genome-wide sequencing approaches that made possible to sequence many individuals, allowing to detect many high reliable polymorphic makers at affordable price (e.g. Schlötterer et al. 2014). For example, pool-sequencing is an approach that consists in sequencing a mixture of genomes of several individuals, instead of sequencing each separately (Turner et al. 2011, Schlötterer et al. 2015, Bailey & Bataillon 2016). This approach allows obtaining allele frequency data, but individual information is lost, preventing linkage disequilibrium analysis. On the other hand, reduced-representation methods such as restriction associated DNA sequencing (RAD-seq), allow obtaining a large amount of individual data but for a smaller number of loci (Davey et al. 2011, Seeb et al. 2014).

Following the advances in sequencing technology, bioinformatic tools have also greatly improved and are essential to deal with the large amount of data generated (Yin et al. 2017). A lot of computational capacity is required to handle the increasing amount of data, databases to store it and efficient and scalable algorithms as well as statistical methods to process and analyze the data. In particular, several softwares have been and continue to be developed, for example, for genome assembly, sequencing alignments, gene finding, SNP identification and genotyping, or genome-wide association studies. Also there are now many available R libraries and python functions that make easier both the tasks of analyzing data and create new programs to deal with specific biological problems.

In the context of this dissertation, I applied several already developed bioinformatic tools, namely for gene ontology analysis (chapter 2), for sequence alignment, for SNP identification and analysis (chapter 3). I have also developed new bioinformatic tools, namely a database for published chromosomal locations of *Drosophila subobscura* genes which I called *DsubSeqLoc* (chapter 2) and a pipeline for removing parental haplotypes (chapter 3).

1.2 Impact of history in adaptive evolution

A relevant topic in evolution studies is the impact of historical differentiation in the outcome of evolution. Historical differentiation (either populations that have become different over time during the study (Blount et al. 2008) or initially differentiated populations when the study starts) may have an impact because random and deterministic processes become interconnected over time, and the

occurrence of future mutations as well as the selective value of existing variants from standing genetic variation may be contingent on the prior history of the population (Jacob et al. 1977, Conte et al. 2012, Lobkovsky & Koonin 2012, Orgogozo 2015).

One evidence of the impact of historical contingencies that arise because similar populations are accumulating different mutations and become somehow differentiated is shown by Blount et al. (2008). Other studies take advantage of initial differentiation between populations, analyzing how much populations converge when evolving under similar conditions. It is expected that 'uniform selection' leads to convergent evolution of laboratory replicated populations and there are several studies that report that (Travisano et al. 1995, Teotónio & Rose, 2000, Spor et al. 2014). Nevertheless this is not always the case, and other studies show that contrasting histories and chance events might prevent convergence from happen. For example, Cohan & Hoffmann (1986, 1989) observed lack of convergence in *Drosophila melanogaster* populations and Plucain et al. (2016) found the same effect for both the growth rate and fitness in *Escherichia coli*. Interestingly, in some cases, convergence is observed at one level but not at the other. Teotónio et al. (2009) observed convergence at phenotypic level, not fully seen at genetic level.

Nevertheless, most studies that report convergent evolution with selection erasing historical signatures were done using lines recently derived from the same ancestral population. This highlights the importance of enlarging the studies to highly differentiated lines, derived from long-term differentiated ancestral populations. This has been the major focus of the studies of *Drosophila subobscura* in the "Local adaptation in *Drosophila*" laboratory where the present study was conducted.

The effects of history of the *Drosophila subobscura* laboratorial populations, derived from natural populations from the extremes of the European latitudinal cline, in their adaptation to a new common environment has been studied by the team where this master thesis was developed, at both the phenotypic (Fragata et al. 2014b, Simões et al. 2017), karyotypic (Fragata et al. 2014a, Simões et al. 2017) and genomic levels (Seabra et al. 2017).

1.2.1. General introduction to *Drosophila subobscura*

Drosophila subobscura is a species of fruit fly that has been the focus of many studies because it presents many chromosomal arrangements (inversions), with latitudinal clinal variation, repeatable across Europe, as well as North and South America (Prevosti et al. 1988, Ayala et al. 1989, Huey et al. 2000, Gilchrist et al. 2004). Several studies in laboratory indicate that this variation is linked to thermal adaptation (Rego et al. 2010, Rezende et al. 2010), though the actual genetic and evolutionary mechanisms are still unknown (e.g. Santos et al. 2005).

The karyotype of *Drosophila subobscura* is composed of five pairs of acrocentric chromosomes (named A, E, J, O and U) and one very small dot. The five pairs of chromosomes mentioned are polymorphic and present latitudinal clines at the genetic level, while the dot chromosome is not polymorphic (Prevosti et al. 1988).

This species is abundant in the Palearctic region, but in 1978 it also appeared in Chile and shortly after spread into Argentina and North America. The source of the colonizers remains uncertain, although all evidence indicates that both the North American and the South American colonizers derived from the same Palearctic population from western Mediterranean or northern European (Ayala et al. 1989).

In a short time period after colonization took place, clines in many chromosomal arrangements evolved in America with identical latitudinal polarity with those found in Europe. This identical polarity in such distant places at the same latitude seems to be a strong evidence that the polymorphisms and the clines are adaptive (Ayala et al. 1989) and are associated with temperature changes (Rego et al. 2010, Rezende et al. 2010). Latitudinal clines in inversion polymorphisms such as these were also found in other species like *Drosophila melanogaster* (Kapun et al. 2016), *Anopheles gambiae* (Fouet et al. 2012) and others. These findings corroborate the adaptive value of inversion polymorphisms, but other factors, like hybridization, founder events, admixture, secondary contact and restricted gene flow may cause the same pattern (Vasemägi 2006, Bergland et al. 2015, Flatt 2016).

Molecular studies in *Drosophila subobscura* also highlight an important impact of chromosomal inversions in genetic patterns of variation (Munté et al. 2005, Simões et al. 2012, Pegueroles et al. 2013, Santos et al. 2016). For example, Simões et al. (2012) found clear genetic differentiation between inversions and also genetic uniformity within chromosomal inversion across a large latitudinal gradient that experiences highly diverse environmental conditions. Santos at al. (2016) studied how the genetic content of inversions evolves during laboratory adaptation, finding evidence of selective changes in the frequency of inversions for seven of 23 chromosomal arrangements, adding further evidence that inversions play a role in adaptation.

The important role of inversions was also shown in other species like *Drosophila pseudoobscura* (Dobzhansky & Epling 1948, Fuller et al. 2016), the silkworm *Bombyx mori* (Ito et al. 2016) or *Anopheles funestus*, the mosquito responsible for malaria disease (Kamdem et al. 2017).

Despite the conclusions of these and many other studies, the mechanisms underlying the evolution of inversions are still not fully understood (Hoffmann and Rieseberg 2008). One important feature of the inversions that is important in their evolution is the reduction of recombination between chromosomes harboring different inversions.

Although there are many genetic studies in *Drosophila subobscura,* including some with the localization of the breakpoints surrounding inversions, there is still no full assembled genome available.

1.2.2. Impact of contrasting genetic backgrounds on evolution at the phenotypic level

To assess the impact of contrasting genetic backgrounds in the adaptation of *Drosophila subobscura* populations to the common environment of the laboratory, Fragata et al. (2014b) studied patterns of phenotypic evolution in populations of *Drosophila subobscura* derived from natural populations located at three European latitudes: one northern (Groningen, Netherlands), one intermediate (Monpellier, France) and one southern (Adraga, Portugal). Collections of flies from the different places were brought to the laboratory founding three-fold replicated laboratorial populations that were studied in real-time ever since.

All populations were maintained with synchronous generations of 28 days, census sizes between 500 and 1200 individuals, photoperiod of 12L:12D and temperature of 18°C . Fragata et al. (2014b) observed high initial differentiation between populations with the Groningen populations having better performance for all life-history traits analyzed, as well as higher starvation resistance and bigger body size.

Throughout generations the fecundity of populations improved considerably and the values presented by each of them became similar. In fact, after only fourteen generations the recently

introduced populations fully converged between them and towards the values of the control (a long established population from Adraga, already in the laboratory for more than 100 generations when the study started). Evolutionary convergence was also seen in starvation resistance, with the Groningen populations, initially with higher values, showing a decrease to the values of both the other recently introduced populations and the control. Finally body size of the flies from Adraga and Montpellier increased to the values of Groningen flies. Overall, then, there was convergence for all phenotypic traits analyzed. Concomitant with this evolutionary convergence, populations showing larger early differentiation relative to the controls presented higher evolutionary rate.

Three years later Simões et al. (2017) conducted another study of real-time laboratory evolution of populations derived from the same locations in Adraga and Groningen, showing that convergent evolution was in general repeatable (predictable) across years, except for body size.

In balance these studies show that in general the sign of history vanishes as time goes by for both life-history and physiological (starvation resistance) traits and that the effect of selection gets stronger with time for life-history traits (further details on Fragata et al. 2014b). Also it is noteworthy that chance events appear to have a bigger role during evolution of starvation resistance than of early fecundity, based on variance components assessed throughout time for these two phenotypic traits.

1.2.3. Impact of contrasting backgrounds at the karyotypic level

Given the high level of polymorphism and the clinal differentiation between *D. subobscura* populations in chromosomal inversions frequencies, a question naturally arise: Do populations converge at the inversion polymorphisms frequencies? Fragata et al. (2014a) addressed this question by analyzing the evolutionary patterns of inversion frequency changes and the impact of evolutionary forces in these changes.

As expected the authors detected high initial differentiation between populations for inversion frequencies at initial generations, reflecting the effect of clinal variation on the geographical origin of the populations. There was variation in the levels of polymorphism between chromosomes, with less variation in the A and J chromosomes. Though the levels of heterozygosity did not differ between populations, allele richness did. Both parameters declined throughout generations.

One important finding was that the populations remained overall differentiated in frequencies of chromosomal inversions, even after 40 generations in the laboratory, in spite the fact that for several inversions selection was involved. Altogether, this indicates that the historical differentiation between foundations at the level of the karyotype had an overall impact on the evolutionary dynamics of inversions in the laboratory.

Simões et al. (2017) also studied at the karyotypic level the populations sampled from the same localities in 2013. They found that initial chromosomal inversion frequencies differed between locations but not between years, meaning that in these three years interval there was no significant differences in inversion frequencies in the sampling from the wild.

All populations exhibited significant changes in inversion frequencies between initial and final generations assayed in the laboratory. After 23/25 generations in the lab, differences between locations remained significant. One interesting difference between studies is that the populations founded in 2013 presented a significant reduction of differences throughout generations, not observed in the populations sampled in 2010 (Simões et al. 2017).

1.2.4. Impact of contrasting backgrounds at the genome-wide level using Pool-seq

After observing convergence of the populations founded in laboratory in 2010 at the phenotypic level but not at the karyotypic level (Fragata et al. 2014a, b) a new question arose: what happened at the genome-wide level in these populations? To answer that question, a genome-wide approach using pool-sequencing methodology was carried out by Seabra et al. (2017), focusing on the Adraga (called Ad) and Groningen (labeled Gro) populations. Samples of each population from four generations (1, 6, 25 and 50, all three-fold replicated except the first) were paired-end sequenced in 4 flow cell lanes of Illumina HiSeq2500 sequencing system, aiming at an average coverage of 50x of each sample. At each generation a synchronous sample pooling the three replicates of the control was also sequenced. In total, thus, 24 samples were pool-sequenced (see Figure 2.1 in Chapter 2).

In that study a total of about 3 million SNPs were obtained and used to follow the evolutionary trajectories of the Adraga and Groningen populations. The number of SNPs in each population decreased during laboratory adaptation, probably indicating that there were many initially rare alleles that got fixed across generations. The nucleotide diversity was similar in both populations in the beginning of the experiment, and it decreased in both populations to values close to the control population.

The two populations were initially differentiated at the genome-wide level (F_{ST} = 0.028) and did not converge, in fact indicating divergence across generations (F_{ST} increased at G50 = 0.042). Interestingly Ad showed a certain degree of convergence to the control, derived from the same geographical origin (F_{ST} decreased from 0.032 to 0.030) but did not fully converge to it.

After analyzing the genome-wide level, SNPs under positive selection were detected in both populations. Interestingly, no common SNPs with signs of selection were found between populations, suggesting different selective responses at the genomic level. Nevertheless there was suggestion that some of the SNPs changed due to selection in both populations. Particularly, there was a peak in allele frequency changes around the candidate SNPs, even in the other population (though being much lower). Also, the differentiation seen around candidate SNPs was also higher in the other population, comparative to random non-candidate SNPs.

Regarding the chromosome location of the candidate SNPs there were major differences between populations. The majority of SNPs under selection between generations 6 and 50 for Ad were located in chromosome E and O while for Gro most were located in chromosome A, the sex chromosome. A Gene Ontology analysis of genes with candidate SNPs was also performed as well as a classification of the type of mutation involved. These last analyses, published in Seabra et al. (2017), constitute part of this dissertation (chapter 2).

1.3 Aims and Dissertation Structure

My master project had two major objectives. The first one was to contribute to the analysis of the genome-wide empirical data, both by Pool-sequencing and RAD-sequencing, to further understand genomic changes occurring during adaptation to the laboratory of two populations of *Drosophila subobscura* from two contrasting latitudes (Adraga, Portugal and Groningen, Netherlands). My second focus was to develop and apply bioinformatic tools to these data. In particular, taking advantage of my previous background in bioinformatics, I contributed to the analysis of the data by developing small programs (R and python scripts), databases and also applying bioinformatics tools to the statistical analysis of the data.

The specific objectives of this thesis were to:

1. Characterize candidate genes associated with SNPs involved in the adaptive process, detected in the analysis of pool-seq data;
2. Develop a database for chromosomal location of genes or genomic regions of *Drosophila subobscura*. This database will be made available and easy to update;
3. Develop a pipeline for the analysis of the sequence data of 4) below;
4. Analyze the RAD-Seq data of individual larvae with known karyotypes, searching for signs of the evolution of genomic content of inversions in general and specific to the adaptive process and how much it differs between populations;
5. Integrate information of SNPs with signs of selection obtained with the Pool-seq and Rad-seq data, in search for common candidate genes/functions involved in the adaptive process as well as in differences between populations.

2. Pool-sequencing

2.1. Introduction

Almost two decades ago, a fully sequenced genome was great news. But population genetic researchers study the population level and not individuals, so one full sequenced genome is not enough, they need more. Also, allele frequencies should be estimated from samples drawn from a large population because the use of small sample sizes can result in considerable errors even when the allele frequencies have been determined at high accuracy (Schlötterer et al. 2014). That is why the arrival of Pool-sequencing (or Pool-seq) is revolutionizing the field.

Pool-seq is the whole-genome sequencing of pools of many individuals and provides a cost-effective alternative to the whole-genome sequencing of individuals, which is very expensive because each individual is sequenced separately, and the preparation of the libraries is also expensive. This latter approach would implicate the sequencing of many individuals in research areas like population genetics. With the increasing availability of new software tools, Pool-seq is being increasingly used for population genomic research on both model and non-model organisms (Schlötterer et al. 2014).

This approach provides genome-wide polymorphism data and this kind of data is becoming increasingly important to serve as a complement to classical genetic analyses. It allows us to know about polymorphic positions in the genome and the frequencies of variant alleles in several populations (Schlötterer et al. 2014). The drawback of this methodology is that haplotypes information is lost which limits linkage disequilibrium analysis.

Seabra et al. (2017) used a Pool-seq approach to study the genome-wide evolution of two *Drosophila subobscura* populations founded in the laboratory from contrasting latitudes, Adraga (Portugal, called Ad) and Groningen (The Netherlands, called Gro) - to understand the genomic mechanisms underlying their laboratory adaptation. Genome resequencing of these populations (three-fold replicated at generation four) was done at four time points since introduction in the laboratory (Figure 2.1). A long-established laboratory control population (TA, also derived from Adraga, in 2001) was also sequenced at the same time points. Candidate SNPs (that gave signs of selection) in Ad and Gro were detected in this study (in both short, G6-G25, and long, G25-G50, periods) and my dissertation work started with the characterization of these SNPs.The objectives were to find a subset of those candidate SNPs with hits with proteins, to make a functional gene ontology characterization to describe biological processes, molecular functions and cellular components affected by those SNPs, and to make a characterization of the types of mutations (synonymous or non-synonymous) behind those SNPs. That part of my master thesis was included in Seabra et al. (2017). Finally, and not included in that paper, I developed a database for *Drosophila subobscura*, that allows to obtain the localization of genes in chromosomes and, if available, locate them within or outside inversions.

There is still no full assembled reference genome available for *Drosophila subobscura* and molecular data on this species is very scarce in databases and datawarehouses available for *Drosophila* genus, like Flybase or Flymine (Gramates et al. 2017 and Lyne et al. 2007 repectively). In spite of this, it is a species studied for many years (for a classic account see Ayala et al. 1989) and there are many papers about this species, including at the population genetic level. The karyotype and the wide number of inversions is characterized, and these inversions are cytologically located (Krimbas 1992, Menozzi and Krimbas 1992, Krimbas 1993, Santos et al. 2005). Also, the sequences of several genes and their cytological location (detected by FISH) are already published but there is no database that

integrates this information. A total of 2250 *Drosophila subobscura* published sequences are available on NCBI and 156 are cytologically located on 20 publications (Appendix 1).

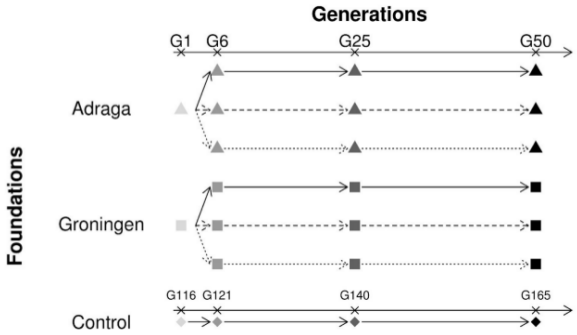

Figure 2.1 – Experimental design of the Pool-seq study. Genome resequencing for pools of 50 individuals from the latitudinal populations (Adraga – triangles, Groningen – squares, Control – diamonds) at four different generations (the generations numbers are marked for each latitudinal population). At generation 1 populations were not yet replicated. After generation 1 replicate populations are marked as: continuous line – replicate 1; dashed line – replicate 2; dotted line – replicate 3. Control populations were sequenced synchronously, pooling all three replicate populations together at each time point analyzed. G1, G6, G25 and G50 correspond to generation 1, 6, 25 and 50 which are the time points analyzed (from Seabra et al. 2017).

2.2. Material and Methods

This dissertation work started with the characterization of some of the candidate SNPs found in the study by Seabra et al. (2017). I analyzed the candidate SNPs (with sign of being under selection) that gave hits with proteins and made a functional gene ontology characterization to describe biological processes, molecular functions and cellular components affected by those SNPs. Also, I made a characterization of the types of mutations (synonymous or non-synonymous) caused by each SNP.

Seabra et al. (2017) assembled a draft reference genome using a homokaryotypic line of this species (*chcu*) (Koske and Maynard Smith 1954) but it is still very fragmented. The draft reference has 2,043 sequences covering 117,329,206 bp, with an average length of 57,429 bp, a maximum length of 820,545 bp and a N50 of 91,130. Assuming a genome size of 120Mb (Adams et al. 2000), 97.5% of the genome is covered by this assembly. This draft reference genome was aligned (BLASTx) against sequences of genes and other genetic regions already published in *Drosophila subobscura* which alowed find published genes in the draft genome assembled.

2.2.1. Gene Ontology analysis

As there is no annotated genome, I followed the following schematically represented in Figure 2.2 to characterize the biological processes, molecular functions and cellular components associated with each protein. I used Gene Ontology framework that provides controlled vocabularies used to describe gene function, and relationships between these concepts.

In a first step, *BLASTx* was used to discover proteins hits in the candidate SNP regions (which are made of 100 bp upstream and 100bp downstream of the candidate SNP, totaling 201 bp region). Only the regions with a match with proteins were further analysed.

In a second step, in order to characterize the mentioned SNP regions, I assessed Gene Ontology categories/terms using *Flybase* and defined higher level categories using QuickGo (http://www.ebi.ac.uk/QuickGO/).

As sometimes more than one term is found to characterize a single gene, decision had to be made to choose the term that best fits the function of the gene. Whenever this was the case, terms inferred from direct assay were chosen over terms inferred from other ways.

Because of the variety and specificity of terms obtained, few proteins were found to be associated with the same process. That way is impossible to find a pattern or direction in the results. To try to solve that problem I tried to group the results in clusters. For that, it was necessary to define higher level categories. This higher level categories, were obtained in the webpage of QuickGo. The QuickGo tools defines relationship between words, in a way that given a word it finds those words whose meaning is related and atributes more general terms as parent terms in the GO graph and more specific terms as child terms.

Some proteins were associated with more than one high level category. That is why in some cases the total number of higher level categories in each dataset is higher than the number of SNPs with hits. Due to the fact that "cellular process" and "multicellular organismal process" terms are very general I did not consider them in this classification.

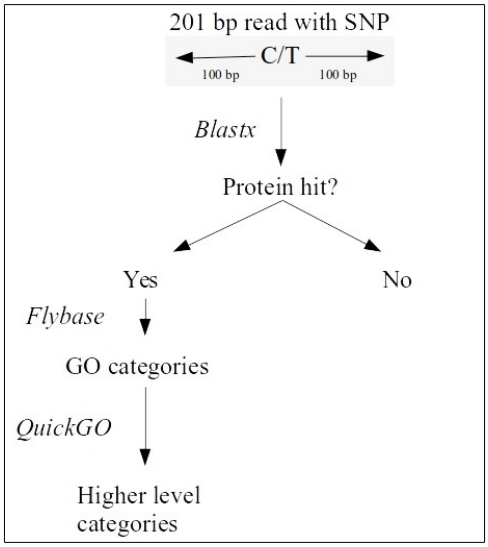

Figure 2.2 –Protocol followed to make the Gene Ontology characterization of the genes.

2.2.2. Type of substitution analysis

To check if the nucleotide substitutions were synonymous or non-synonymous I used the methodology described in Figure 2.3. For the sequences with hits with proteins (the same as in 2.2.1) I searched for a the nucleotide sequence in the nt database (*blastn*) (accessed at 24/3/2017). The parameters used were the default ones and it were selected the sequences with higher score values in the alignment to be used as reference. To discover the type of mutation I aligned each pair of sequences (201bp sequence with the SNP and the sequence with highest score resulting from Blastn alignment) and made the translation of the two sequences using *BioEdit version 7.1.9* (Hall 1999).

This translation allowed to discover the three nucleotide codon generated by each sequenced and the amino acid codified was search on codon table in http://www.chegg.com.

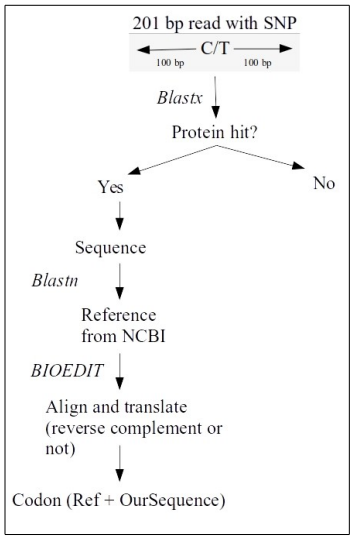

Figure 2.3 – Protocol followed to make the classification of the type of mutation in each SNP region.

2.2.3. *DsubSeqLoc* Database

Flybase, a database of *Drosophila* Genes & Genomes, only presents information about other species like *Drosophila melanogaster* or the American counterpart of *Drosophila subobscura*, *Drosophila pseudoobscura* (the closer species phylogenetically). Flymine is a datawarehouse that presents information about many *Drosophila* species, including *D. subobscura* (namely published papers about this species), but does not include cytological information.

In order to assemble the information already published about sequences of genes and other regions of *D. subobscura*, as well as their cytological location, and to relate them with the new data of the draft genome that we developed in our laboratory (Seabra et al. 2017), I developed a database called *DsubSeqLoc*. It was constructed in *phpmyadmin 4.6.5.2* using *mySQL 2.0* and is composed of several entities and relation tables, namely gene names, reference to papers that published nucleotide sequences of those genes, and the cytological location of those genes, including, whenever available, their location in relation to inversions. To relate information published with the new data of the draft genome it was done the *BLAST* of genes in the chromosomes of *Drosophila subobscura* against the draft reference genome.

To make easier the process of searching the database and to make it available to any user, even those who lack knowledge on how to make queries to the database or dealing with mySQL, I created a website that gives access to the content of the database. The website was created in PHP and css using the *https://html5up.net/read-only* free template. In the index.php file I wrote code to create search boxes and connect them to the database. The connection to the database was done using mySQLi PHP extention (Gilmore 2006).

2.3. Results

2.3.1. Characterization of SNPs

From the 134 and 288 candidate SNPs indicating selection between generations 6 and 25 (G6-G25) for Ad and Gro, respectively, I observed significant hits with proteins in 24 and 36 SNP regions, respectively. For the sets of SNPs indicating selection between generations 6 and 50 (G6-G50), I found hits with proteins for 37 of the 189 candidate SNPs for Adraga, and for 13 of the 107 candidate SNPs for Groningen. A varied number of biological processes are involved, with the most represented higher level GO categories, present in all datasets, being localization, metabolic processes and response to stimulus Others, such as biological regulation and rhythmic processes, were only present in Adraga (Figure 2.4).

Although the low number of hits with proteins precludes a proper comparative analysis between Ad and Gro populations, it is anyway tempting to discuss the processes that have been detected as being under selection. The new laboratory environment subjects the flies to new conditions such as density, both in juveniles and adults, age of reproduction, temperature, nutrients, photoperiod, among others (Pegueroles et al. 1999). It is thus no surprise that in general metabolic processes are affected (anabolism and catabolism), as well as processes involved in the distribution of molecules (localization), responses to stimuli and rhythmic processes, e.g. because time exposition to light has changed.

The only common family of genes harboring SNPs detected to be under selection both in Adraga and Groningen was a gustatory receptor gene family, but these SNPs are located in different chromosomes: gustatory receptor 22a and 22d at chromosome U for Ad and gustatory receptor 59e at chromosome E for Gro.

A few of the identified genes had more than one SNP significant for selection (1 gene for Adraga and 4 genes for Groningen). There was a total of 12 scaffolds with two or three genes harboring candidate SNPs. None of the genes with candidate SNPs were located in common scaffolds between Adraga and Groningen.

Bottom line, I did not detect convergent evolution between Adraga and Groningen populations at the gene/protein levels of organization. Nevertheless, as said before, the few hits with proteins, and the different number of proteins with hits of SNPs under selection for Ad and Gro do not allow a proper analysis and comparison to fully explore this scenario.

I found a lower number of synonymous than non-synonymous substitutions in Adraga, whereas in Groningen the number of non-synonymous and synonymous substitutions was similar (Table 2.1). Interestingly, some of the mutations under selection were located in small intronic regions in all datasets.

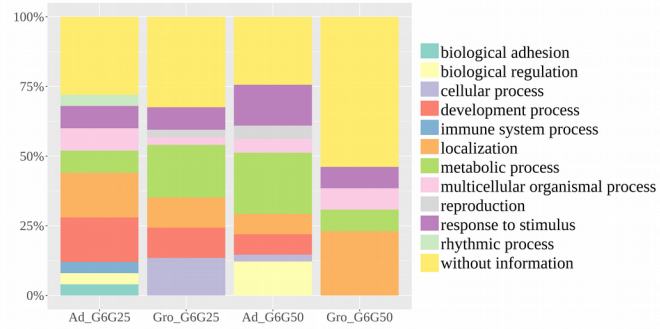

Figure 2.4 – Bar charts with the higher level GO categories found for candidate SNPs (with significant protein hits) detected in Ad and Gro in both short and long periods.

Table 2.1 – Number of candidate SNPs with significant protein hits and from these, those that are synonymous, non-synonymous or that are located in introns, for both Ad and Gro in short and long period.

	Ad G6-G25	Ad G6-G50	Gro G6-G25	Gro G6-G50
Total	134	189	288	107
Protein hits	24	37	36	13
Non-synonymous	14	22	15	2
Synonymous	5	4	13	3
Intron	2	7	5	5

2.3.2. Database

The database structure utilized to store and retrieve information was modelled as a relational database (Figure 2.5). This database is composed of 9 tables (5 corresponding to entities and 4 corresponding to relationships) due to the underlying nature of selected data. It stores 20 links to papers and sequence id of the sequences in NCBI. The webpage that gives easier access to that database is available at the link http://www-personal.fc.ul.pt/~mmmatos/DsubSeqLoc.

The database websearch allows the user to introduce a name of a gene and obtain information on its location (chromosome and cytological band). Also, when available it provides information whether the gene is inside or outside given chromosomal inversions.

The database also allows the user to insert, not only a single gene but also a list of genes and obtain the corresponding location information for each gene in the list. I used the database to locate the candidate SNPs obtained by pool-sequencing inside or outside inversions.

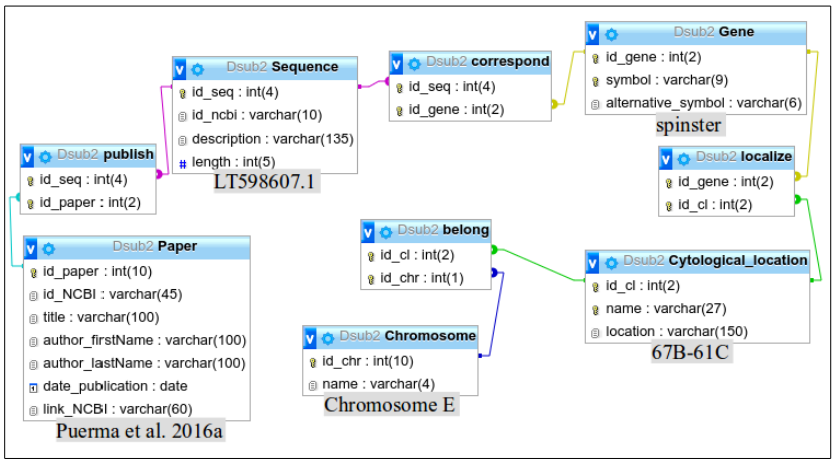

Figure 2.5 – Entity Relationship (ER) model of the *DsubSeqLoc* database constructed. Below each table, in gray, it is shown an example of information within the database. The example correspond to a gene matching a candidate SNP in Ad detected in Pool-seq analysis of long period adaptation (G6G50).

3. RAD-sequencing

3.1. Introduction

Whenever we need to have individual genetic information for many individuals, whole genome sequencing is still very expensive. This limitation was the booster for the arrival of a method for genome-wide genetic markers development and genotyping called Restriction site-associated DNA sequencing or, in short, RAD-seq (Baird et al. 2008, Emerson et al. 2010, Hohenlohe et al. 2010, Davey & Blaxter 2011, Davey et al. 2011).

This method uses restriction enzyme digestion of target genomes to reduce the complexity of the target and it has shown to be capable of delivering thousands of sequenced markers across many individuals for any organism at reasonable costs. This method has the advantage, relatively to pool-seq, of preserving the identity of individuals because it uses molecular identifiers (MID) to associate sequence reads to particular individuals. Also, it allows to have high reliability on the called SNPs if coverage is high while the reliability of the ones captured by pool-seq is only moderate (Davey & Blaxter 2011, Schlötterer et al. 2014).

In short, the method includes five major steps: cut of individual DNA by the restriction enzyme; the fragments are ligated to a P1 adapter that contains a sticky end that makes the ligation possible to happen and a MID (molecular identifier that will uniquely identify the individual and tag the fragment); the fragments are pooled and then sheared to generate shorter fragments; all the fragments will be ligated to a P2 adapter and, in the end, will be amplified using two primers. Only the fragments with both P1 and P2 adapters will be amplified because of the characteristics of the P2 adapter (Davey & Blaxter 2011) (Figure 3.1).

In this work I aim to understand the evolutionary dynamics of the genomic content of chromosomes with specific inversions and how they differ between Ad and Gro populations of *Drosophila subobscura*. For that, sequencing separately individuals with known karyotypes, was required. RAD-sequencing was a natural choice for this objective.

Studying the genomic content of chromosomes with specific inversions is important because, in spite the role of these inversions in processes such as adaptation, speciation and the evolution of sex chromosomes, the underlying evolutionary mechanisms are not fully understood. In particular, there is a high controversy in what concerns the evolution of inversions and the processes that are involved in the maintenance of inversion polymorphisms in natural populations. While the coadaptation hypothesis (Dobzhanky 1950, Dobzhanky 1970) is based on a selective advantage of inversion heterokaryotypes due to the existence of positive epistatic interactions between loci located within chromosomal arrangements, the local adaptation hypothesis (Kirkpatrick and Barton 2006) states that as long as chromosomal inversions present sets of alleles adapted to local conditions, they may be selected even without epistasis. The spread of an inversion can be thus explained by the maintenance of a given set of alleles with positive effects on fitness. While it is not easy to disentangle the two hypotheses, the study of the genomic content of inversions, how much it differs between populations and how it evolves during adaptive evolution may shed light on these issues (Hoffmann & Rieseberg 2008, Simões et al. 2012, Fragata et al. 2014a, Santos et al. 2016).

History and selection are likely to shape the evolution of inversions but at what extent it is not known. Fragata el al. 2014a report signs of positive selection for some inversions, but they were

variable between populations, that maintained differentiation for inversions after 40 generations of laboratory evolution (Fragata et al. 2014a).

In this chapter I present the analysis that I carried out in this thesis of data obtained by RAD-sequencing of a large number of individual larvae of *Drosophila subobscura* from replicate populations coming from Adraga, Portugal and Groningen, Netherlands at two generations, 6 and 25, of evolution in the laboratory. The chromosomal inversions of these larvae were previously characterized through cytological analysis (Fragata et al. 2014a). The same populations and generations were analyzed at the genome-wide level by pool-sequencing by Seabra et al. (2017) and in this thesis (Chapter 2) which allowed to make a comparative analysis between studies.

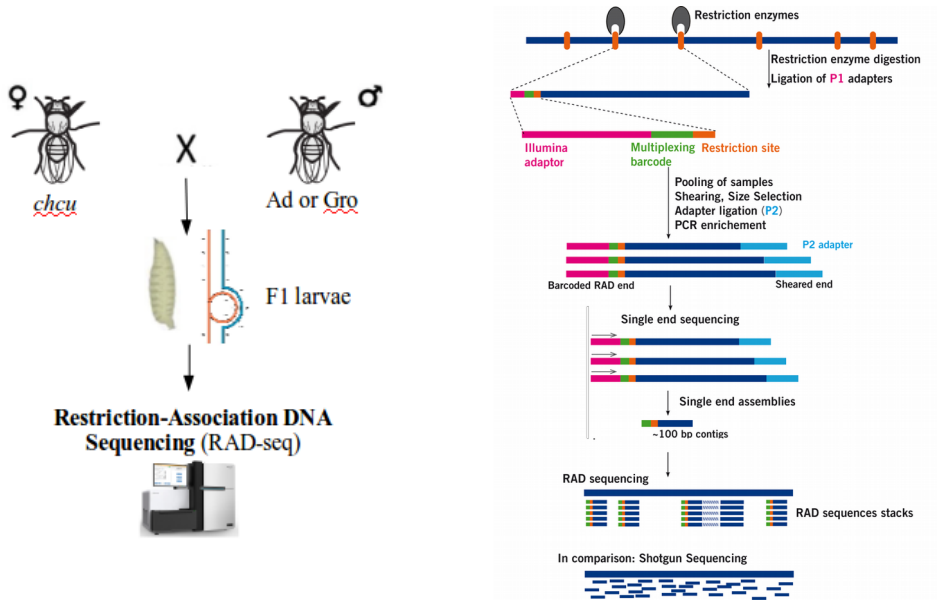

Figure 3.1 – Schematic representation of the crosses that allowed to obtain the F1 larvae that were sequenced (on the left) and RAD-Sequencing protocol (on the right) (Adapted from Floragenex Technical Brief)

3.2. Material and Methods

3.2.1. Biological material and RAD-sequencing

This part of the work was done previously to my arrival in the laboratory, but I present here this information since it is not published yet and is important to understand the subsequent flow of analysis.

Two populations of *Drosophila subobscura* were collected from two contrasting latitudes (Adraga, Portugal and Groningen, Netherlands), brought to a new common environment, the laboratory environment. These populations were threefold-replicated at the fourth generation after founding (details in Fragata et al. 2014a). RAD-sequencing was done on individuals from three replicates of each population (Ad1, Ad2, Ad3, Gro1, Gro2 and Gro3) from generation 6 (G6) and 25 (G25) after introduction in the laboratory.

To analyze the genomic content of chromosomes wild males (whose chromosomes do not have recombination) from the populations were crossed with females of the homokaryotypic lineage *chcu* (Figure 3.1). This cross was done to allow the characterization of the chromosomal arrangements. *Chcu* is a isogenic, homokaryotypic strain, with the following chromosomal arrangements: A_{ST} , J_{ST} ,

U_{ST} , E_{ST} and O_{3+4} (Balanyà et al. 2004). The cross generates F1 larvae that have half the genome of the *chcu* lineage and half the genome of the wild individual. The cytological visualization of the chromosomes of the F1 larvae allows to identify the arrangement of each paternal, that is "wild", chromosome due to the formation of specific loops in the alignment of homologous chromosomes (details in Simões et al. 2012). RAD-sequencing was done in a set of the larvae in the corresponding populations and generations (see details below). To ensure that less frequent, but still interesting, chromosomal arrangements were sequenced, after a random choice of the individual larvae, some others were also included, to a final number of c. 39 individuals per replicate population and generation.

Given that half the genome of each individual was *chcu*, to be able to remove the corresponding haplotypes from the sequences we also sequenced *chcu* individuals. In total 475 larvaes were sequenced, 117 from Ad at G6, 117 from Gro at G6, 115 from Ad at G25, 115 from Gro at G25 and 11 *chcu* (see Appendix 2).

DNA from each individual larva was extracted using the phenol-chloroform extraction method (Sambrook and Russell 2001). Individual DNA extracts were distributed in five 96-well plates for sequencing, and it was essential to equalize the DNA amount and concentration on each plate (5 ng/ul in 13 ul total for each individual). Precise quantification was done with the Qubit 2.0 Fluorometer (Invitrogen). DNA extracted was sent to Floragenex (http://www.floragenex.com/) that prepared the RAD libraries using PstI restriction enzyme and carried out single-end 100 bp sequencing in Illumina HiSeq2500 platform, using one sequencing lane per each of the 96-sample plates, sequenced twice to double the sequencing amount for a good coverage.

From the single-end Illumina sequencing of the 475 individuals, we obtained an average of 3.3 M reads per individual. The Fastq files obtained from sequencing were checked for base quality in *FastQC*. Each fastq file included reads from several individuals, identified by sequence barcodes. Reads were processed by *process_radtags* (from the package *Stacks* (Catchen et al. 2013)) to remove reads with uncalled bases and with low quality scores, to check that the barcode and restriction site are intact in each read and to demultiplex the samples based on the barcode identification. After the later mentioned processing and filtering steps, an average of 2.9 M reads per individual were retained.

3.2.2. RAD-seq analysis

The software Stacks was chosen to process and analyse RAD-seq data. This software is composed of several components that can be used separately. The first one is the already mentioned *process_radtags* which process reads. When a reference genome is available, as is our case (Seabra et al. 2017), the reads are aligned to this reference using and alignment program *bowtie2 version 2.2.1* (Langmead et al. 2009). Then, the second component of *Stacks,* named *pstacks,* extracts stacks (RAD loci) that have been aligned to the reference and identifies SNPs. The minimum depth of coverage allowed to report a stack was 3. The third component is *cstacks* and assembles the catalog based on alignment position, not sequence identity. The fourth component, *sstacks* makes matches each individual reads to the catalog.

After running these components, a fifth one, *populations,* allows exporting loci and SNP data after applying filters for missing data. This program was executed for three sets of data separately: one consisting of the 11 individuals of *chcu*, another consisting in all 232 individuals of Ad and another of 232 individuals of Gro. The filters applied to obtain reliable loci in all 3 cases were: at least half of the individuals in a population must have data on a locus to process that locus for that population (-r 50)

and a minimum stack depth of 10 is required for individuals at that locus (-m 10). For two sets of data (Ad and Gro), there was the additional filter that a locus must be present in at least 3 of the 12 samples (Ad1G6, Ad2G6, Ad3G6, Ad1G25, Ad2G25, Ad3G25, Gro1G6, Gro2G6, Gro3G6, Gro1G25, Gro2G25 and Gro3G25) .

3.2.3. Pipeline for removal of *chcu* haplotypes

From each individual I needed to remove the haplotype from *chcu*, to keep only the haplotype from the population I am studying. For that purpose I developed a pipeline that consists of a workflow of programs mainly written in *python 2.7* programming language and linux shell to process datafiles in sequence.

This pipeline is composed of 12 programs, the main program is called *RH_pipeline.py* that import other python programs created by me. When necessary, this main program uses *subprocess* module to spawn linux shell processes. This pipeline is an output of this thesis and all the steps are detailed in the results (section 3.3.2).

3.2.4. Statistical analysis of RAD-seq data

In the previous step I obtained, for each individual and for each polymorphic site (SNP), the allele coming from our "wild" population (Ad or Gro), after excluding the allele coming from *chcu*. Since most software of SNP analysis require diploid codification I duplicated the allele, obtaining a total matrix of 417264 SNPs for 462 individuals in map and ped format.

There is no complete reference genome for this species and in our fragmented draft genome (see chapter 2, section 2.2) we have no information about the location of the chromosomal arrangements (inversions). However, we know to which chromosome each of our fragments (scaffolds) belong, from homology with *D. melanogaster*. Thus, I was able to analyze the SNPs located in each of the 5 chromosomes separately (see section 3.3.3). Since I know which inversion was present in each individual, I was also able to analyze the SNPs present in each chromosome with a specific chromosomal arrangement. I selected chromosomal arrangements O_{3+4}, O_{ST} and A_2 (see sections 3.3.4.1, 3.3.4.2 and 3.3.5.1) for several reasons. Two of them (O_{3+4} and O_{ST}) were chosen because they are located in the chromosome O, the one with more molecular information available. Also, the O_{3+4} arrangement presents an interesting dynamic, as it increases in frequency in Ad but not in Gro. The A_2 inversion is interesting to analyze because it is present on the sexual chromosome, occurs with high frequency in both populations, allowing good sample size to study. This inversion presents a temporal increase in frequency in both populations. From these arrangements, I was interested in finding those SNPs with signs of selection, and as a surrogate of cytogenomic location, I also analyzed SNPs with statistical association with each inversion.

I performed a Principal Component Analysis (PCA) to visually assess the genome-wide differentiation between individual samples. The Principal Component Analysis was done with a R script that uses the libraries *SNPRelate* v1.6.4 and *gdsfmt* (Zheng et al. 2012) from *bioconductor* development software project. I used the *snpgdsVCF2bGDS* function to reformat Variant Call Format (VCF) file and the *snpgdsPCA* function to calculate the eigenvectors and eigenvalues for principal component analysis.

To assess differentiation between populations and across generations, I estimated pairwise mean F_{ST} (Weir and Cockerham 1984) between groups (replicates/populations/generations) using the program *vcftools version 3.0* (Danecek et al. 2011) with --weir-fst-pop argument and plotted a

Principal Coordinates Analysis (PCoA) to visually assess this differentiation. To plot the PCoA I ran a R script that computes principal coordinate decomposition with the function *pcoa* from package *ape version 4.1* (Paradis et al. 2004).

Among the set of SNPs located in chromosomes with a given chromosomal arrangement, I searched for SNPs with signs of selection, using a conservative approach. I wanted to keep SNPs that change by selection forces and avoid those that may change due to genetic drift. With this goal in mind, I used the two first conditions for detecting SNPs with signs of selection that Seabra et al. (2017) used: 1) To find those SNPs with highest significant values in Cochran-Mantel-Haenszel (CMH) test (Mantel and Haenszel 1959) and 2) From these, to find those SNPs whose minor allele increased in all three replicates (SNPs with only one allele at G6 were not considered in selection test). To compensate the lack of step 3 in Seabra et al (2017), specifically simulations to finally disentangle changes between generations expected by drift alone, the cut-line of the CMH –log p value here defined was in general higher than in Seabra et al (2017) – see appendix 3. In any case, because I did not include the final step of Seabra et al (2017), the comparison of results of the two studies needs to be made with care.

For each population and generation, I performed an association analysis between each candidate SNP and its related arrangement (that is individuals with versus without that arrangement), using all data of the three replicates not discriminated. The analysis also allowed to detect if the number of SNPs associated with a given inversion will increase or decrease throughout time. The significance of the association was estimated by Fisher's exact test after FDR correction (adjusted P value for alpha = 0.05) (Benjamini and Yekutieli 2001, theorem 1.3). I used to option –fisher of *p-link* 1.9 (v1.07).

3.2.5. Linkage disequilibrium analysis

The RAD-seq approach has an important advantage which is to allow linkage disequilibrium analysis. As I do not have a full assembled reference genome I assessed linkage disequilibrium patterns for SNPs located in a given scaffold. I applied this analysis to chromosomes of G6 with O_{3+4} arrangement in each scaffold. This arrangement serves here as a mere illustration, but others will certainly be analysed later.

The linkage disequilibrium was calculated as the mean of the linkage values in the three replicates (1, 2 and 3) per population (Ad or Gro) using –geno-r2 argument from *vcftools version 3.0* (Danecek et al. 2011).

In the context of this work I compare the same scaffolds in the two populations to assess whether these two populations present similar patterns of linkage in the same scaffolds.

3.3 Results

3.3.1. Assessing missing data and distribution of SNPs per locus

I analyzed the data of 462 individuals of our experimental populations, 232 of Ad and 230 of Gro, characterizing 89.092 loci for Ad and 89.996 loci for Gro.

To perform the RAD-seq analysis, first of all I made a characterization of my data. This first step is really important as the choice on the filtering values will influence all further analyses. With this goal, I plotted the distribution of missing data per individual (Figure 3.2) and per locus (Figure

3.3) for both Ad and Gro populations using two programs created by me using the generic R function hist (Becker et al. 1988, Venables and Ripley 2002). The R program *plot_distribution_missingData_per_locus* receives as input *haplotypes.tsv* file (each line corresponds to one haplotype file) created by *Stacks* and counts the missing data in each line. The R program *plot_distribution_missingData_per_ind* receives the same file but transposed (each line represents one individual) and counts the missing data per line. The majority of the individual presents no information for 10.000 to 20.000 locus both in Ad (11.2% to 22.5% Figure 3.2A) and Gro (11.1% to 22.2% Figure 3.2B). In terms of missing data per locus most loci have information for all individuals, or at least for most of them (Figure 3.3). The number of SNPs per locus was, on average, 8.8 in Ad and 8.9 in Gro.

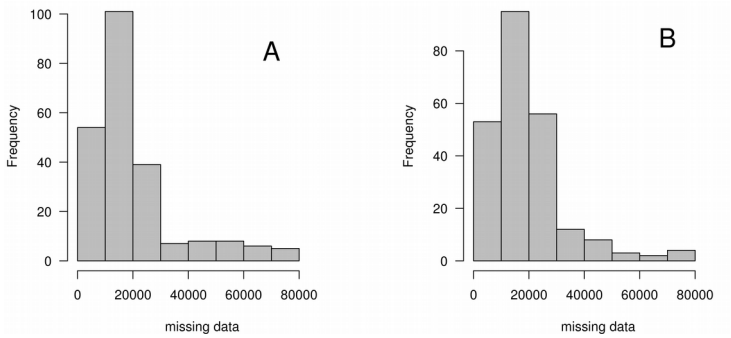

Figure 3.2 – Histograms of distribution of number of loci with missing data per individual A) in Ad and B) in Gro.

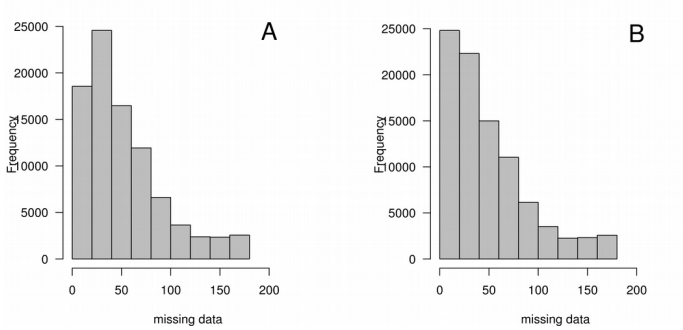

Figure 3.3 – Histograms of distribution of number of individuals with missing data per locus A) in Ad and B) in Gro.

3.3.2. Pipeline

The pipeline developed (Figure 3.4) (https://github.com/marta-antunes/remove-haplotypes) receives as input three files generated by *Stacks*, one with haplotypes of *chcu* (an hapstats.tsv file generated by *populations* program from *Stacks*), another with the larvae sequences (including thus both the populations haplotypes and *chcu* haplotypes) and another that allow me to extract the position of the SNPs (catalog.tags).

In this pipeline, a dictionary of positions is created. This dictionary allows to include the correct information about the position of each SNP, since haplotypes file does not contain information about the position of each SNP, but the position corresponding to the beginning of the tag. Another

dicionary is filled with haplotypes from *chcu*. Then, a python program iterates over haplotypes.tsv file and if the haplotype is in the *chcu* dictionary, that haplotype is removed. The pipeline includes also a filtering step that allows the user to choose the percentage of missing data allowed. The pipeline generates four important files as output: the plink flat files (map and ped), a file with frequencies and a file that keeps records of filtering removals.

In this work, I used the pipeline to remove *chcu* haplotypes and generate the corresponding map and ped files. I chose the allowed percentage of missing data to be 25 percent, meaning that if a locus misses information in more than 25 percent of individuals that locus is discarded. I provide a usage example: python RH_pipeline.py /pop_Ad/batch_1.haplotypes.tsv pop_Ad/batch_1.hapstats.tsv/ pop_chcu/batch_1.hapstats.tsv 25. In this example the first file corresponds to file with Ad haplotypes, the second corresponds to the file with correct positions of the Ad SNPs and the third file correspond to the file with *chcu* haplotypes. The last argument in this example corresponds to the percentage of missing data allowed.

As stated above, the pipeline that I developed was used in this work to remove *chcu* haplotypes from the RAD-seq data, but has other applications. It allows to remove parental haplotypes from hybrid species, that is an application similar to the one presented in this dissertation but it also can be used to remove a desirable haplotype from individuals, even if they are not parent and progeny.

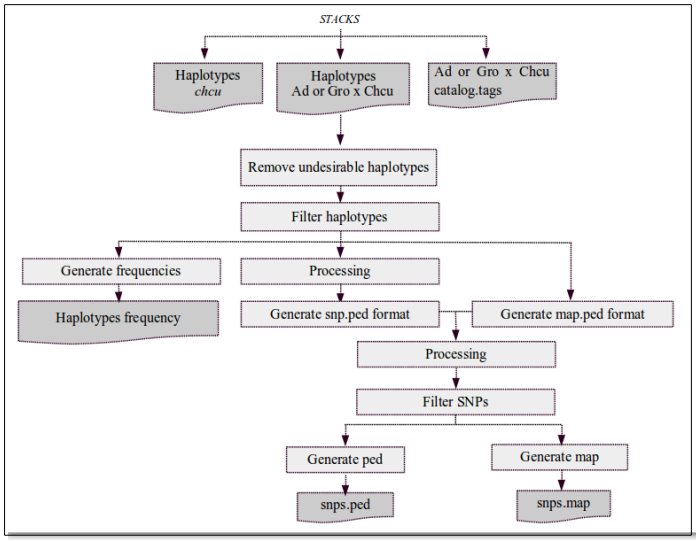

Figure 3.4 – Remove *chcu* haplotypes pipeline scheme.

3.3.3. Analyses of RAD-sequencing data

In total, there were 417264 SNPs in the whole dataset. I detected some differentiation between Ad and Gro populations and also between generations but this latter is less pronounced (Figure 3.5A). The differentiation between G6 and G25 is more evident when we look at the populations separately (Figure 3.5B and 10C) and is higher in Ad.

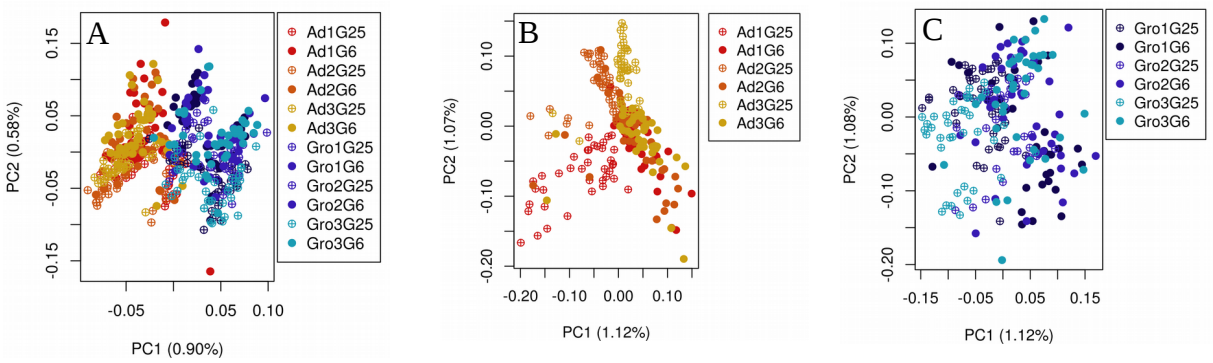

Figure 3.5 – Principal Component Analysis of SNP variation at the genome-wide level. A) for all SNPs, B) for SNPs in Ad and C) for SNPs in Gro.

As one of the aims of this work is to understand the evolutionary dynamics of the genomic content of chromosomes, I separated the analysis of the data per chromosome and studied the genome-wide differentiation in each of the five chromosomes of *Drosophila subobscura* (Figure 3.6). I found that the individuals are more clearly separated by the inversions they carry than by the population to which they belong. This is observed for all chromosomes.

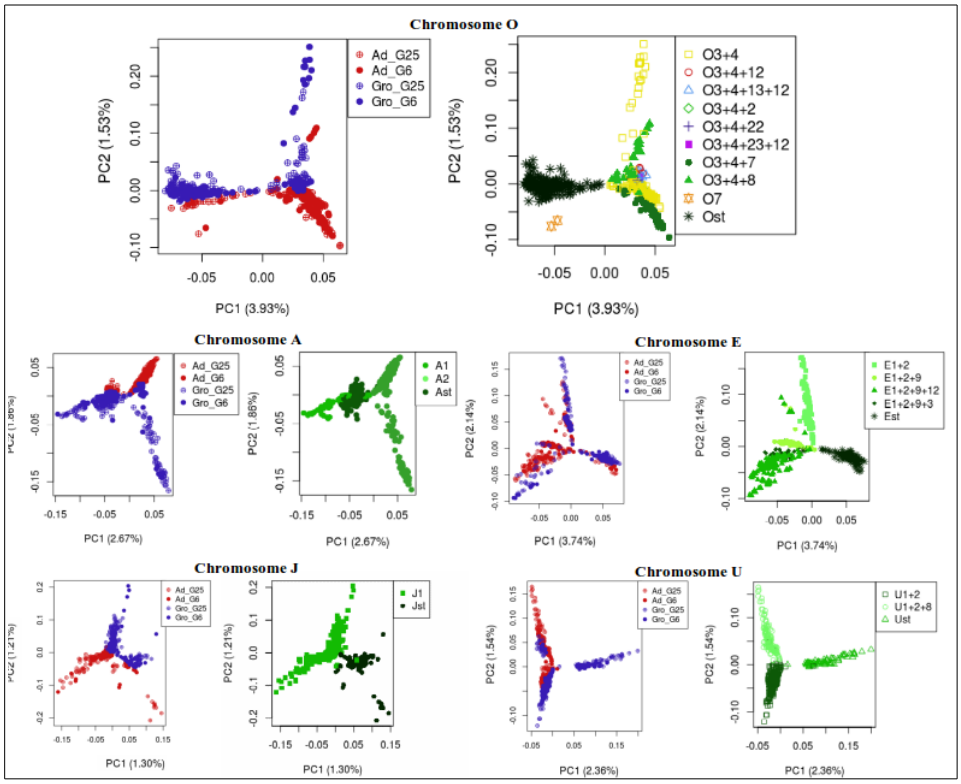

Figure 3.6 – Principal Component Analysis of SNP variation in each of the five chromosomes. This figure is composed of five pairs of plots. The plot on the left in each pair correspond to the Principal component analysis of SNP variation highlighting populations information (Ad_G6 - red filled circle; Ad_G25 - red crossed circle; Gro_G6 - blue filled circle; Gro_G25 - blue crossed circle). The plot on the right in each pair correspond to the Principal component analysis of SNP variation highlighting inversion information (different gradients of green correspond to different inversions or arrangements).

3.3.4. Analysis of chromosome O

3.3.4.1. Analysis of chromosomes with O_{3+4} arrangement

Chromosome O is the one with more information available at the molecular level in part because of the available Va/Ba (Varicose/Bare) balancer stocks (Sperlich et al. 1977, Santos 2009), allowing the relatively easy generation of isogenic, homokaryotypic O lines from wild O chromosomes. I analyzed the genomic changes occurring in this chromosome, focusing on two of the chromosomal arrangements. I analyzed the data in the O chromosome of individuals with the O_{3+4} arrangement as well as the O_{ST} arrangement (see this below). Here I will detail the analysis of the O_{3+4} arrangement. In total there are 157 individuals with this arrangement and 40210 SNPs.

The differentiation in individuals with O_{3+4} arrangement was analyzed at two levels: temporal (differentiation between generations in each population) and dynamic of differentiation between poulations (how the differentiation between Ad and Gro changed across time). The PCA suggests some differentiation between generations in Ad (Figure 3.7) and this differentiation is confirmed by the F_{ST} analysis (mean F_{ST} = 0.044, Table 3.1). The differentiation between generations is not so clear in Gro individuals (Table 3.1) and the F_{ST} values involving samples of generation 25 of Gro1 or Gro3 are all negative. This negative values at Gro1G25 and Gro3G25 are meaningless because there is only one individual Gro1 from generation 25 and also only one individual Gro3 at the same generation with arrangement O_{3+4} and this lack of data bias the results. Thus only the Gro2 values may allow some interpretation.

In terms of dynamic, although at G6 the individuals from Ad and Gro populations do not appear to be much differentiated in the PCA plot (Figure 3.7), the F_{ST} values indicate that the populations are somehow differentiated, though not much (average 0,028 between all Ad replicates and all Gro replicates, negative values were considered zero). The differentiation between Ad and Gro increases from G6 to G25, with an increase of 0.02, 0.03 and 0.08 when comparing Ad1, Ad2, Ad3 with Gro2 respectively (Table 3.1).

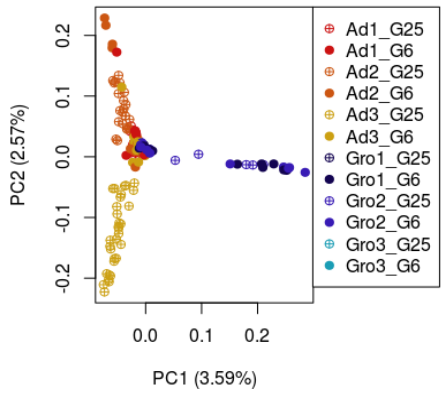

Figure 3.7 – PCA of SNP variation in chromosomes with O_{3+4} arrangement.

Table 3.1 – Matrix of mean pairwise F_{ST} between groups of individuals for SNPs located in chromosomes with O_{3+4} arrangement (40210 SNPs). In pink: between generations of the same replicate population; in blue: between populations at generation 6; and in green: between populations at generation 25.

	Ad1_G6	Ad1_G25	Ad2_G6	Ad2_G25	Ad3_G6	Ad3_G25	Gro1_G6	Gro1_G25	Gro2_G6	Gro2_G25	Gro3_G6	Gro3_G25
Ad1_G6	0.000	0.016	0.012	0.043	0.003	0.067	0.035	-0.594	0.033	0.028	-0.031	-0.567
Ad1_G25		0.000	0.036	0.062	0.018	0.085	0.052	-0.509	0.053	0.058	0.008	-0.478
Ad2_G6			0.000	0.048	0.020	0.084	0.054	-0.505	0.054	0.060	0.007	-0.479
Ad2_G25				0.000	0.045	0.089	0.077	-0.473	0.079	0.087	0.040	-0.455
Ad3_G6					0.000	0.069	0.035	-0.603	0.033	0.029	-0.029	-0.561
Ad3_G25						0.000	0.097	-0.455	0.103	0.115	0.076	-0.423
Gro1_G6							0.000	-0.521	-0.012	-0.028	-0.055	-0.491
Gro1_G25								0.000	-0.490	-0.345	-0.396	nan
Gro2_G6									0.000	-0.026	-0.060	-0.468
Gro2_G25										0.000	-0.035	-0.295
Gro3_G6											0.000	-0.368
Gro3_G25												0.000

SNPs under selection

I tested for SNPs under positive selection in individuals with O_{3+4} arrangement and detected 36 candidate SNPs in Ad. There is a clear separation between generations for these SNPs, particularly in Ad, as expected (Figure 3.8A). F_{ST} differentiation values between generations were 0.109, 0.211 and 0.339 for the replicates of Ad (Ad1, Ad2 and Ad3 respectively) (Table 3.2) validating the differentiation observed in PCA. The same SNPs (the ones that are under selection in Ad) were analyzed in Gro, and none indicated changes consistent with selection, this is, there is no differentiation between generations for these SNPs in Gro. Concomitantly, in terms of dynamics, populations Ad and Gro did not converge for the candidate SNPs of Ad (F_{ST} between Ad and Gro increases from 0.039 in G6 to 0.046 in G25, 0.046 to 0.095 and 0.018 to 0.212 when comparing Ad1, Ad2, Ad3 with Gro2 respectively, Table 3.2).

I also tested for SNPs under positive selection in Gro, detecting only 2 candidate SNPs in this population (Figure 3.8B). Gro2 individuals are separated in these SNPs between generations (F_{ST} between the two Gro2 samples was 0.159, Table 3.2) but Ad samples did not show differentiation between generations (negative values of F_{ST}, Table 3.2).

In terms of dynamics, populations Ad and Gro appear to converge for the candidate SNPs of Gro because differentiation between Ad and Gro appears to be decreasing between generations. Nevertheless this result can be biased by the small number of SNPs analyzed, and, most importantly, by the fact that only Gro2 replicate is being considered in the analysis.

In fact, as there is only one individual Gro1 and one individual Gro3 at G25, the estimated frequencies of alleles in these samples will always be either 1 or 0 (depending on the presence or absence of that allele in the individual). This way the frequencies will not represent the actual frequencies in the population. This will have implications in the number of candidate SNPs detected because of the "minor allele increase in all replicates" condition imposed on my data. This condition makes the SNPs for which the allele that was minor at G6 and that is not present in the Gro1 or Gro2 G25 individuals to be discarded. This may result in the detection of a smaller number of candidate SNPs, as the 2 SNPs that I obtained.

I detected candidate SNPs for selection in Gro a second time, but this time, I used only the replicated population that has enough number of individuals to calculate correct frequencies (Gro2) and detected minor alleles increasing between generations of this replicate. Following this

methodology I obtained 69 SNPs but this number of SNPS does not implicate the important condition of consistency across replicates. Having this in consideration, I observed that, in terms of dynamics, the pattern obtained is different from the one shown by the 2 SNPs (Figure 3.9 and Figure 3.8B respectively). This second approach indicates that SNPs are responding to selection in this population Gro but not, or in less extent in Ad. This appears to indicate no convergence for these SNPs, but again these SNPs are not increasing in all replicates (Table 3.3).

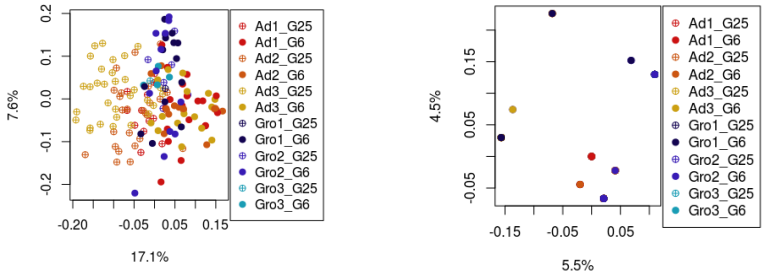

Figure 3.8 – PCA of SNPs under selection variation A) in Ad and B) in Gro of individuals with O_{3+4} arrangement.

Table 3.2 – Matrix of mean pairwise F_{ST} between groups of individuals for SNPs located in chromosomes with O_{3+4} arrangement and that show signs of selection in Ad (above the main diagonal; 36 SNPs) and in Gro (below the main diagnonal; 2 SNPs). In pink: between generations of the same replicate population; in blue: between populations at generation 6; and in green: between populations at generation 25.

36 SNPs / 2 SNPs	Ad1_G6	Ad1_G25	Ad2_G6	Ad2_G25	Ad3_G6	Ad3_G25	Gro1_G6	Gro1_G25	Gro2_G6	Gro2_G25	Gro3_G6	Gro3_G25
Ad1_G6	0.000	0.109	-0.015	0.218	-0.023	0.339	0.065	-0.416	0.039	0.022	0.061	-0.175
Ad1_G25	-0.056	0.000	0.144	0.057	0.132	0.133	0.047	0.040	0.016	0.046	-0.056	-0.263
Ad2_G6	-0.048	-0.100	0.000	0.211	-0.024	0.344	0.061	-0.710	0.046	0.033	0.021	-0.221
Ad2_G25	-0.040	-0.037	-0.054	0.000	0.205	0.092	0.118	-0.003	0.085	0.095	0.032	-0.160
Ad3_G6	-0.018	-0.010	-0.051	-0.051	0.000	0.339	0.053	-0.516	0.018	0.016	0.004	-0.118
Ad3_G25	0.096	0.120	0.056	0.029	-0.026	0.000	0.196	-0.030	0.170	0.212	0.121	-0.038
Gro1_G6	0.122	0.160	0.203	0.214	0.256	0.385	0.000	-0.433	-0.045	-0.064	-0.072	-0.256
Gro1_G25	0.000	0.020	-0.100	-0.122	-0.267	-0.393	0.518	0.000	-0.349	-0.337	-0.458	-nan
Gro2_G6	0.029	0.024	0.068	0.125	0.175	0.334	-0.046	0.400	0.000	-0.060	-0.088	-0.225
Gro2_G25	-0.092	-0.020	-0.082	-0.145	-0.206	-0.208	0.202	0.000	0.159	0.000	-0.047	-0.390
Gro3_G6	-0.093	-0.010	0.041	0.009	0.067	0.235	-0.121	0.000	-0.139	-0.069	0.000	-0.214
Gro3_G25	0.000	0.020	-0.100	-0.122	-0.267	-0.393	0.518	-nan	0.400	0.000	0.000	0.000

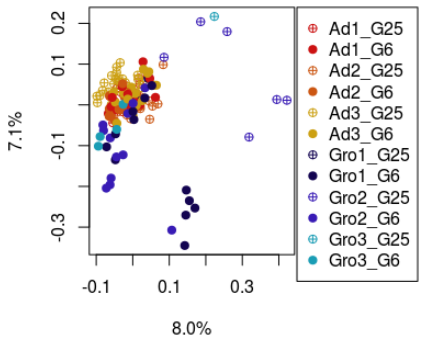

Figure 3.9 – PCA SNPs under selection (increase in just 1 replicate) in Gro of individuals with O_{3+4} arrangement.

Table 3.3 – Matrix of mean pairwise F_{ST} between groups of individuals for SNPs located in chromosomes with O_{3+4} arrangement and that show signs of selection in Gro (frequency of the minor allele increasing in just one replicate). In pink are shown F_{ST} values between generations, in blue F_{ST} values between poulations at generation 6 and in green F_{ST} values between populations at generation 25.

69SNPs	Ad1_G6	Ad1_G25	Ad2_G6	Ad2_G25	Ad3_G6	Ad3_G25	Gro1_G6	Gro1_G25	Gro2_G6	Gro2_G25	Gro3_G6	Gro3_G25
Ad1_G6	0.000	0.001	0.017	0.046	-0.005	0.066	0.053	0.015	0.068	0.260	0.017	-0.014
Ad1_G25		0.000	0.036	0.073	0.006	0.052	0.075	0.158	0.087	0.284	0.048	0.166
Ad2_G6			0.000	0.076	0.039	0.070	0.094	0.099	0.065	0.310	0.065	0.123
Ad2_G25				0.000	0.053	0.130	0.088	-0.052	0.110	0.280	0.093	0.066
Ad3_G6					0.000	0.061	0.060	0.069	0.088	0.196	0.012	-0.044
Ad3_G25						0.000	0.137	0.136	0.133	0.282	0.070	0.045
Gro1_G6							0.000	0.226	0.013	0.275	0.013	0.116
Gro1_G25								0.000	0.243	0.073	0.689	-nan
Gro2_G6									0.000	0.421	-0.076	0.133
Gro2_G25										0.000	0.385	-0.275
Gro3_G6											0.000	0.410
Gro3_G25												0.000

Association analysis

From the SNPs with signs of selection in Ad from the previous analysis, I detected 15 SNPs (42%) associated with the O_{3+4} chromosomal arrangement in G6 and 24 (67%) in G25. The number of SNPs associated with the inversion thus increases from G6 to G25 in Ad. This makes sense because the frequency of this inversion is increasing in Ad population (Simões et al. 2017).

In Gro I detected no SNPs (of the 2 significant for selection) associated with the inversion in G6 and 1 (50%) in G25. The number of SNPs associated with the inversion increases from G6 to G25 in Gro, although the frequency of this inversion is decreasing in this population when adapting to the laboratory environment. But this result is meaningless given the number of SNPs in this analysis.

Linkage disequilibrium analysis

Although I have calculated the linkage disequilibrium for all scaffolds of G6 chromosomes with O_{3+4} arrangement, here I only present the patterns observed in two scaffolds which correspond to scaffolds where SNPs under selection in Ad were found and also have been located in relation to the arrangement (Figure 3.10).

In general, I detected variable patterns of linkage desiquilibrium within each scaffold. Interestingly the patterns of linkage appear to be higher outside the O_{3+4} arrangement than the ones found for the scaffold within the arrangement (Figure 3.1). However, in other scaffolds I found higher linkage values inside the arrangement (not show). Also, average values found do not indicate higher LD outside than inside the arrangement (see below).

The average linkage disequilibrium in scaffold outside the arrangement was not very different between populations (mean R^2 in Ad is 0.14 and in Gro is 0.11). However, the correlation of LD across the scaffold between the two populations was low (0.3). For the scaffold within arrangement, similar average values of linkage were detected in both Ad and Gro (0.13). Importantly, the correlation was of LD between populations was very low (0.04).

Given that, in both cases, there was a low correlation between LD of Ad and Gro, this could indicate a different genomic content associated with the contrasting linkage patterns between populations in these scaffolds. This is particularly seen for the scaffold within the arrangement which may be expected due to the low recombination in that region.

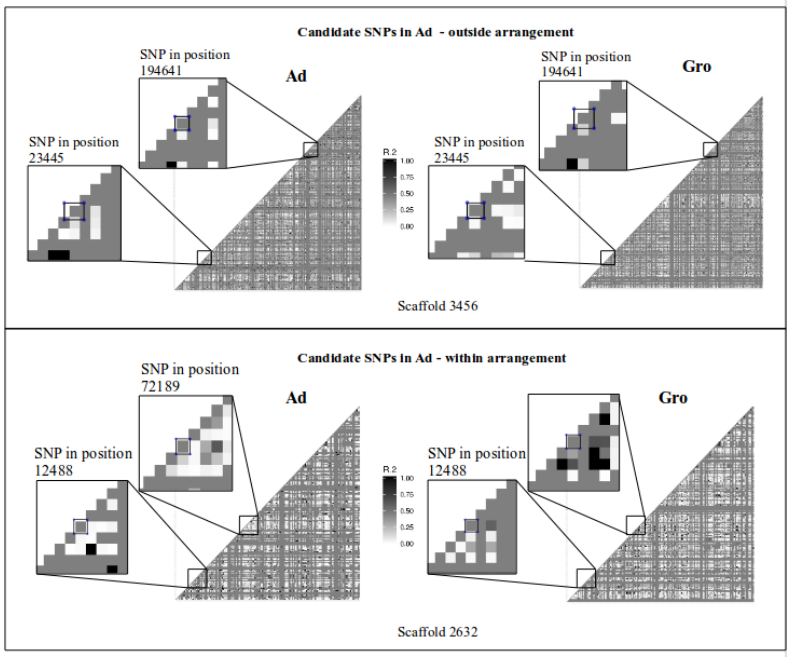

Figure 3.10 – Linkage Disequilibrium (R^2) heat maps in two scaffolds harboring SNPs with signs of selection in Ad. Top: scaffold located outside O_{3+4} arrangement; bottom: scaffold located inside arrangement; left: in Ad; right in Gro.

3.3.4.2. Analysis of chromosomes with O_{ST} inversion

The PCA plot of variation of SNPs located in chromosomes with O_{ST} arrangement shows differentiation between generations of Gro but not between generations of Ad (Figure 3.11). In accordance, the F_{ST} values indicate higher differentiation between generations of Gro (F_{ST} values of 0.024, 0.016 and 0.021 for Gro1, Gro2 and Gro3 respectively) than between generations of Ad (Table 3.4). In fact, two of the F_{ST} values between generations of Ad (Ad1 and Ad3) are meaningless because there is only one individual Ad1 from generation 6 and one individual Ad3 from generation 25 with arrangement O_{ST} and this lack of data bias the results. Thus only the Ad2 values may allow some interpretation. Nevertheless, this value for Ad2 is negative which indicates that there is no differentiation between generations of Ad.

There is no differentiation between Ad and Gro in SNPs located in chromosomes with O_{ST} inversion at G6, although this is not a very strong comparison because there are only a total of 5 individuals of Ad sequenced at this generation (1 Ad1, 2 Ad2 and 2 Ad3).

In terms of dynamic, the F_{ST} values indicate that differentiation between Ad and Gro increases from G6 to G25, with a mean increase of 0.067 when comparing all 3 replicates of Gro with Ad1 and all three replicates of Gro with Ad2 and assuming negative F_{ST} values as no differentiation between Ad and Gro (Table 3.4).

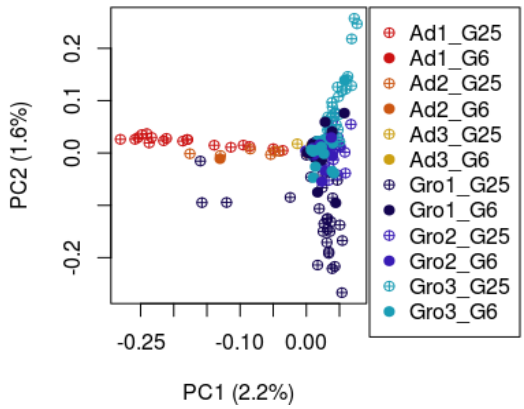

Figure 3.11 – PCA of individuals with O_{ST} inversion.

Table 3.4 – Matrix of mean pairwise FST between groups of individuals for SNPs located in chromosomes with O_{ST} inversion (40210 SNPs). In pink: between generations of the same replicate population; in blue: between populations at generation 6; and in green: between populations at generation 25.

	Ad1_G6	Ad1_G25	Ad2_G6	Ad2_G25	Ad3_G6	Ad3_G25	Gro1_G6	Gro1_G25	Gro2_G6	Gro2_G25	Gro3_G6	Gro3_G25
Ad1_G6	0.000	-0.309	-0.305	-0.264	-0.319	nan	-0.577	-0.538	-0.576	-0.543	-0.610	-0.531
Ad1_G25		0.000	-0.053	0.117	0.019	-0.296	0.072	0.085	0.072	0.086	0.063	0.090
Ad2_G6			0.000	-0.127	-0.129	-0.291	-0.240	-0.211	-0.241	-0.210	-0.260	-0.201
Ad2_G25				0.000	-0.002	-0.266	0.020	0.046	0.022	0.045	0.015	0.051
Ad3_G6					0.000	-0.366	-0.190	-0.153	-0.191	-0.160	-0.209	-0.154
Ad3_G25						0.000	-0.573	-0.527	-0.574	-0.538	-0.605	-0.528
Gro1_G6							0.000	0.024	0.000	0.020	-0.003	0.022
Gro1_G25								0.000	0.025	0.037	0.026	0.046
Gro2_G6									0.000	0.016	-0.001	0.024
Gro2_G25										0.000	0.016	0.031
Gro3_G6											0.000	0.021
Gro3_G25												0.000

SNPs under selection

Since I excluded in the conditions to define signs of selection, SNPs that were at least for one population, fixed at G6, there are no SNPs under selection in Ad: for Ad1 naturally having only data of one individual in generation 6 causes an apparent fixation; also Ad3 has only one individual in generation 25, which causes a bias in the calculation of allele frequencies at generation 25.

64 SNPs gave signs of selection in Gro. The individuals are clearly separated in these SNPs between generations 6 and 25 in Gro (Figure 3.12) with F_{ST} values of 0.177, 0.120 and 0.177, respectively for each replicate (Table 3.5). In Ad the differentiation between generations can only be analyzed in one replicate (Ad2) with F_{ST} value for this replicate of 0.036 (Table 3.5). This indicates that the SNPs under selection in Gro are responding in Gro and also could be responding in Ad, based on the observation of the differentiation between generations of Ad2, that contrasts with no differentiation estimated with the entire set of SNPs (Table 3.4).

In terms of dynamic, populations Ad and Gro did not converge for the candidate SNPs in Gro. In fact, differentiation between the two populations in general increased from G6 to G25, with a mean

increase of 0.159 comparing all replicates of Gro with Ad1 and Ad2 (Table 3.5). This increase is higher than the one observed for the entire set of SNPs, suggesting that, in contrast with what might be expected, populations diverge more for SNPs under selection than for SNPs under drift. This is an interesting finding, one that is in accordance with the finding of Seabra et al. (2017) with the pool-seq analysis of the same populations.

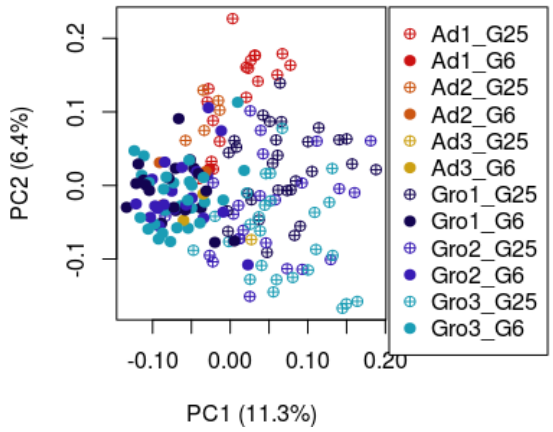

Figure 3.12 – PCA SNPs under selection in Gro of individuals with O_{ST} inversion.

Table 3.5 – Matrix of mean pairwise FST between groups of individuals for SNPs located in chromosomes with O_{ST} inversion and that show signs of selection in Gro (64 SNPs). In pink: between generations of the same replicate population; in blue: between populations at generation 6; and in green: between populations at generation 25.

	Ad1_G6	Ad1_G25	Ad2_G6	Ad2_G25	Ad3_G6	Ad3_G25	Gro1_G6	Gro1_G25	Gro2_G6	Gro2_G25	Gro3_G6	Gro3_G25
Ad1_G6	0.000	-0.014	-0.619	-0.062	-0.286	nan	-0.320	-0.132	-0.239	-0.065	-0.267	-0.039
Ad1_G25		0.000	0.162	0.191	0.228	0.182	0.177	0.161	0.180	0.197	0.201	0.218
Ad2_G6			0.000	0.036	-0.189	-0.385	-0.264	-0.044	-0.225	-0.039	-0.236	-0.013
Ad2_G25				0.000	0.143	-0.004	0.104	0.115	0.068	0.109	0.074	0.156
Ad3_G6					0.000	-0.125	-0.164	0.011	-0.161	0.003	-0.133	-0.008
Ad3_G25						0.000	-0.241	-0.074	-0.231	-0.143	-0.250	-0.169
Gro1_G6							0.000	0.177	-0.016	0.165	-0.010	0.187
Gro1_G25								0.000	0.161	0.061	0.188	0.088
Gro2_G6									0.000	0.120	-0.015	0.159
Gro2_G25										0.000	0.144	0.032
Gro3_G6											0.000	0.177
Gro3_G25												0.000

Association Analysis

Of the set of SNPs with signs of selection I detected 36 (56%) associated with the inversion in GroG6 and 44 (69%) in GroG25. The number of SNPs associated with the inversion increases from G6 to G25. This makes sense because the frequency of this inversion is also increasing (Simões et al. 2017).

3.3.5 Analysis of chromosome A

3.3.5.1. Analysis of Chromosomes with inversion A_2

In total there are 282 individuals with A_2 inversion. PCA and PCoA revealed some differentiation between generations in both Ad and Gro, higher in Gro, for SNPs located in chromosomes with A_2 arrangement (mean F_{ST} between generations is 0.026 in Ad and 0.079 in Gro) (Figure 3.13 and 19, Table 3.6).

Differentiation between Ad and Gro in G6 was on average 0.017 (Table 3.6, Figure 24). The differentiation between populations increases, on average being 0.101 (mean F_{ST} value at G25), having a minimum increase of 0.06 and a maximum increase of 0.1 (see Table 3.6, Figure 3.14).

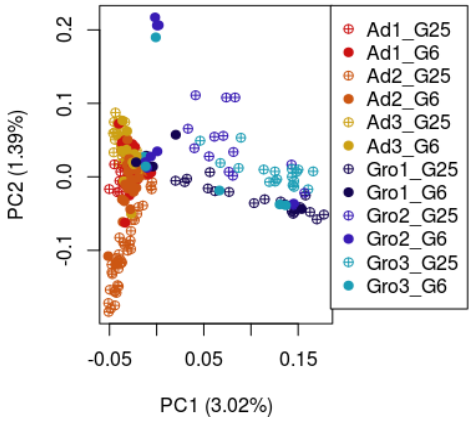

Figure 3.13 – PCA of individuals with A_2 inversion.

Table 3.6 – Matrix of mean pairwise F_{ST} between groups of individuals for SNPs located in chromosomes with A_2 inversion (24921 SNPs). In pink: between generations of the same replicate population; in blue: between populations at generation 6; and in green: between populations at generation 25.

	Ad1_G6	Ad1_G25	Ad2_G6	Ad2_G25	Ad3_G6	Ad3_G25	Gro1_G6	Gro1_G25	Gro2_G6	Gro2_G25	Gro3_G6	Gro3_G25
Ad1_G6	0.000	0.021	0.005	0.034	0.006	0.027	0.004	0.073	0.021	0.051	0.017	0.084
Ad1_G25		0.000	0.028	0.055	0.029	0.047	0.036	0.100	0.054	0.080	0.050	0.115
Ad2_G6			0.000	0.028	0.009	0.030	0.007	0.076	0.024	0.054	0.020	0.086
Ad2_G25				0.000	0.038	0.054	0.052	0.107	0.072	0.091	0.067	0.121
Ad3_G6					0.000	0.028	0.010	0.074	0.025	0.053	0.021	0.084
Ad3_G25						0.000	0.041	0.099	0.059	0.082	0.054	0.111
Gro1_G6							0.000	0.081	0.007	0.036	-0.002	0.104
Gro1_G25								0.000	0.105	0.059	0.069	0.081
Gro2_G6									0.000	0.049	-0.002	0.133
Gro2_G25										0.000	0.038	0.057
Gro3_G6											0.000	0.108
Gro3_G25												0.000

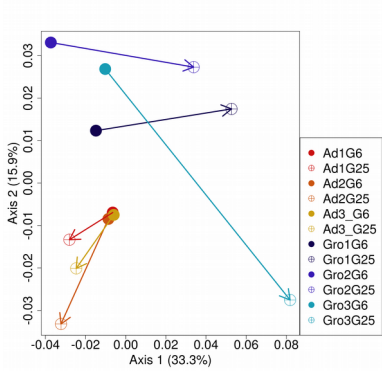

Figure 3.14 – PCoA using F_{ST} of individuals with A_2 inversion.

SNPs under selection

56 SNPs were detected as being under selection in Ad and 100 in Gro. Considering the candidate SNPs in Ad, the PCA suggests a clear differentiation between generations of Ad individuals, as expected, but not of Gro individuals (Figure 3.15A). In spite of this, the F_{ST} analysis indicates also some differentiation between generations in Gro population, although smaller (mean F_{ST} in Ad is 0.136, mean F_{ST} in Gro is 0.113) (Table 3.7). Importantly, F_{ST} between generations in Gro is bigger than for the entire set of SNPs, suggesting that though not indicating selection these SNPs do show a higher dynamic than the one expected by drift alone. Populations do not converge for the candidate SNPs in Ad as differentiation between Ad and Gro increases in all replicates (Table 3.7, Figure 3.16A). In fact the increase of differentiation of Ad and Gro between G6 and G25 is higher (F_{ST}G25-F_{ST}G6 = 0.105) than for the whole set of SNPs (F_{ST}G25-F_{ST}G6 = 0.084).

Analyzing the candidate SNPs in Gro, the PCA indicates a clear differentiation between generations of Gro individuals but not of Ad individuals (Figure 3.15B). As expected, F_{ST} analysis shows higher differentiation between generations in Gro. Though smaller, some differentiation is also seen in Ad (mean F_{ST} Gro = 0.389, mean F_{ST} Ad = 0.036) (Table 3.7). Though not much higher, the F_{ST} between generations in Ad is bigger than for the entire set of SNPs, suggesting, as previously for Gro (considering the SNPs under selection of Ad), that these SNPs have a higher dynamic than for SNPs under drift alone. Also, as seen for the candidate SNPs of Ad, Ad and Gro populations also do not converge for the candidate SNPs in Gro (Table 3.7, Figure 3.16B), in fact they present a divergence across generations, with a differentiation increasing of, on average, 0.472 higher than for the whole set of SNPs.

Two SNPs were common to Ad and Gro: one located in scaffold 6157 in position 6470 and the other in scaffold2735 in position 454. These SNPs were searched on the *DsubSeqLoc* database created in the context of this master project and it was not possible to located them yet.

The F_{ST} values between generations are higher for SNPs under selection than for all SNPs in chromosome A. This is observed for both candidate SNPs in Ad and in Gro.

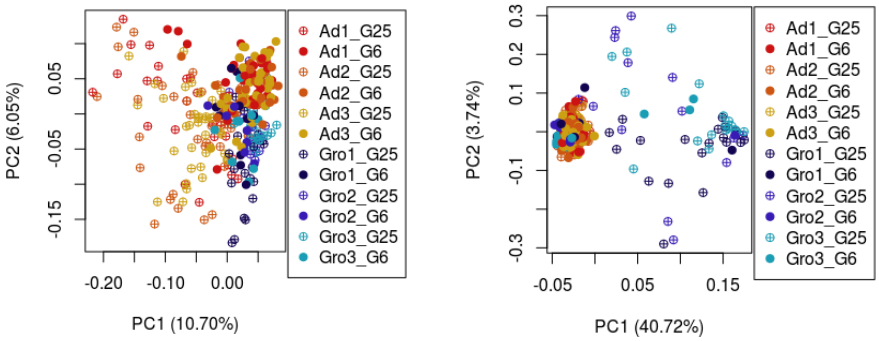

Figure 3.15 – PCA of individuals with A_2 inversion A) SNPs under selection Ad and B) SNPs under selection Gro.

Table 3.7 –Matrix of mean pairwise F_{ST} between groups of individuals for SNPs located in chromosomes with A_2 inversion and that show signs of selection in Ad (above the main diagonal; 56 SNPs) and in Gro (below the main diagnonal; 100 SNPs). In pink: between generations of the same replicate population; in blue: between populations at generation 6; and in green: between populations at generation 25.

100 SNPs \ 56 SNPs	Ad1_G6	Ad1_G25	Ad2_G6	Ad2_G25	Ad3_G6	Ad3_G25	Gro1_G6	Gro1_G25	Gro2_G6	Gro2_G25	Gro3_G6	Gro3_G25
Ad1_G6	0.000	0.098	-0.008	0.146	-0.005	0.122	0.040	0.153	0.060	0.072	0.046	0.116
Ad1_G25	0.040	0.000	0.126	0.036	0.128	0.017	0.066	0.157	0.064	0.134	0.044	0.183
Ad2_G6	0.000	0.042	0.000	0.163	-0.002	0.145	0.046	0.132	0.068	0.065	0.055	0.095
Ad2_G25	0.037	0.058	0.032	0.000	0.171	0.039	0.096	0.157	0.080	0.135	0.093	0.191
Ad3_G6	0.002	0.044	0.008	0.050	0.000	0.148	0.056	0.133	0.071	0.069	0.043	0.095
Ad3_G25	0.028	0.054	0.034	0.046	0.036	0.000	0.081	0.140	0.070	0.137	0.080	0.196
Gro1_G6	0.033	0.091	0.047	0.107	0.053	0.067	0.000	0.128	0.017	0.081	-0.022	0.150
Gro1_G25	0.582	0.573	0.586	0.602	0.574	0.585	0.506	0.000	0.111	0.077	0.055	0.125
Gro2_G6	0.030	0.082	0.041	0.101	0.044	0.074	-0.102	0.501	0.000	0.062	0.009	0.171
Gro2_G25	0.384	0.400	0.393	0.423	0.379	0.403	0.233	0.106	0.219	0.000	0.060	0.115
Gro3_G6	0.095	0.147	0.113	0.176	0.101	0.134	-0.068	0.421	-0.103	0.147	0.000	0.150
Gro3_G25	0.599	0.597	0.603	0.619	0.591	0.604	0.526	0.039	0.526	0.091	0.443	0.000

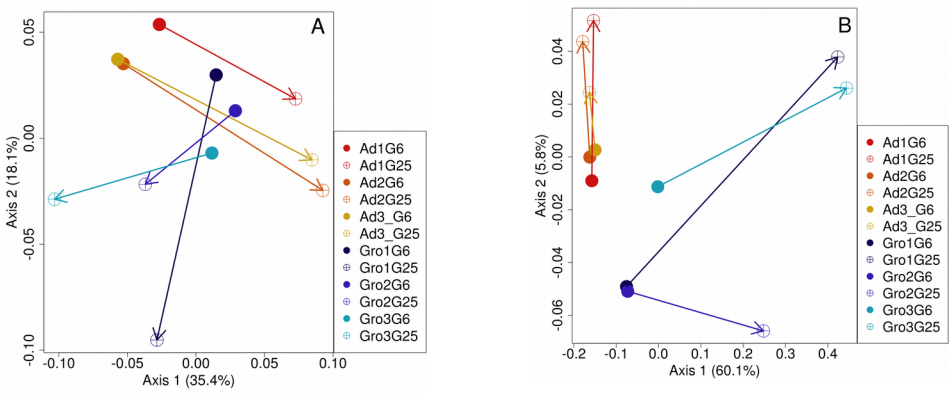

Figure 3.16 – PCoA of individuals with A_2 inversion A) SNPs under selection Ad and B) SNPs under selection Gro.

The low number of candidate SNPs detected in the analyses done in this work can be partly due to the fact that one SNP that gives signs of selection in one population can be fixed in the other. The differences can be caused merely by the existence of several variants and not because differences in genetic background affects the selective pressures that change the direction of evolution. In fact, in chromosomes with A2 inversion 41% of the SNPs that give signs of selection in Ad are fixed in the Gro population and 36% of the SNPs that give signs of selection in Gro are fixed in the Ad population. Thus the differences observed in the sets of SNPs indicating selection in Ad and Gro were in fact partly due to this lack of common genetic variation.

Association analysis

In Ad I detected 10 SNPs (18%) associated with the inversion in G6 and 24 (43%) in G25. In Gro I detected 57 SNPs (57%) associated with the inversion in G6 and 93 (93%) in G25.

It was previously shown that, in both Ad and Gro there is an increase in the frequency of the A2 inversion during laboratory adaptation, bigger in Ad though also present in Gro. This indicates that this inversion may have a higher adaptive value relative to the others, in particular Ast. In this work, I observed an increase in the number of candidate SNPs associated with the inversion over time in both Ad and Gro. This suggestion that selection increases the differences in the genetic content of the inversions is interesting and deserves future analysis.

4. General Discussion

4.1. Main Achievements

With this dissertation, I made a preliminary analysis of the biological processes and types of mutations underlying the evolutionary changes at the genomic level of populations of *Drosophila subobscura* as they adapt to new, laboratorial conditions. These populations were derived from contrasting European latitudes (Adraga, Portugal and Groningen, Netherlands) and were pool-sequenced at several generations after laboratory introduction (chapter 2, Seabra et al. 2017). Moreover, to contribute to the understanding of the evolutionary and genetic mechanisms involved in chromosomal inversions evolution, I analyzed by RAD-sequencing many individuals with known karyotypes from the same populations and generations (chapter 3). With this thesis I also increased the amount of bioinformatic tools available to analyze not only *Drosophila subobscura* data but also data from others species (chapter 2 and 3). Below I will discuss the results of each part of the study.

I developed a database, *DsubSeqLoc*, that integrates information that was available on nucleotide sequences and cytological location of genes or of other genomic regions in *Drosophila subobscura*. The creation of this database allows easier access to this information. I also developed a pipeline that allows to remove one of the parental haplotypes from the progeny. This pipeline can be applied not only to our *Drosophila subobscura* data, but to any other case in which we know the genotype(s) of one of the parents.

In the pool-sequencing study (chapter 2), integrated in an 'evolve and resequence' study involving pool-sequencing (Seabra et al. 2017), I have made Gene Ontology (GO) analysis of proteins in regions of SNPs with signs of selection (chapter 2, Seabra et al. 2017). This analysis reinforces the overall conclusion that the genomic evolution does not lead to convergence between populations because many different biological processes, that differ between populations, respond to selection during laboratory adaptation of *Drosophila subobscura* populations (Fragata et al. 2014b, Seabra et al. 2017). This suggests a polygenic basis in the studied traits (Barton & Keightley 2002, Seabra et al. 2017) and highlights the importance of the historical genetic background in the evolutionary responses of populations. A scenario of rugged fitness landscapes (Wright 1932, de Visser and Krug 2014) may be involved here as convergence was observed at the phenotypic level (Fragata et al. 2014b) but not at the genomic level (Seabra et al. 2017), suggesting different genetic paths to reach the same phenotypic outcome (fitness). Several studies show that convergent molecular changes are more common at gene level than at the nucleotide level (Tenaillon et al. 2012, Dettman et al. 2012, Orgogozo 2015,). Despite that, we did not find genes under selection in common between the two populations except for one family of genes. This may be a reinforcement of the theory that different pathways are been taking to achieve the same outcome.

Curiously, during my search for the types of mutations of the candidate SNPs, I found that some SNPs were located in short intronic regions, which are reported to be abundant in the genome of *Drosophila* (Parsch et al. 2010). The fact that short intronic regions are under selection may be due to the occurrence of alternative splicing. Actually, (Farlow et al. 2012) showed that the signature of selection is stonger on shorter introns, which present weaker splice sites, than in longer introns (see Farlow et al. 2012). Finally, we cannot exclude the role of linkage disequilibrium between a SNP indicating selection and the real target of selection (hitchhiking), for these and, in general, all candidate SNPs (Tobler et al. 2014).

In the RAD-sequencing study (chapter 3), we found that the individuals differ more by the inversions they have than by the populations to which they belong. This shows, for the first time at a genome-wide scale in *D. subobscura*, that inversions are much differentiated, even within the same population. This same pattern was seen previously with microsatellites (Simões et al. 2012).

We then analyzed the entire set of SNPs located in a chromosome carrying a specific inversion (O_{3+4}, O_{ST} and A_2). At the beginning of the experiment, populations were differentiated at chromosomes carrying O_{3+4} and A_2, but not at those carrying O_{ST} inversion. The question remains whether this differentiation is caused by SNPs outside and/or inside the inversions. The lack of population differentiation in chromosomes carrying O_{ST} may be due to small sample size, as only 5 Ad individuals bearing this inversion were sequenced. But we cannot exclude the role of high gene flow between populations bearing this inversion, as reported by Pegueroles et al. (2013).

The RAD-seq data also revealed no genetic convergence between populations, either at the genome-wide level or for candidate SNPs. On the contrary, our data indicates that populations diverged between generations. Importantly, they even diverged more for candidate SNPs than for genome-wide (Appendix 5). This is in accordance with what was found by Seabra el al. (2017) in the genome-wide pool-seq study for the same populations. In other words, in common with that study with poolseq data, here with the Rad-seq analysis I observed that history prevented genomic convergence to happen (Cohan and Hoffmann 1986, Cohan and Hoffmann 1989, Plucain et al. 2016) and that populations explore different genetic pathways of the adaptive landscape to reach the same final state (Wright 1932, de Visser and Krug 2014).

In spite the general disparities between populations in the sets of candidate SNPs detected, it is important to note that some of them may be actually responding to selection in all populations, even in populations where they were not detected, e.g. because of smaller changes of frequencies of the selected allele due to higher values since the starting generations. In fact, we found that the differentiation across generations in one population for the candidate SNPs of the other population was bigger than for the whole set of SNPs (Appendix 4). Again this is in accordance with Seabra et al. (2017).

One important goal in the analysis of our RAD-seq data is to contribute to clarify controversies about which mechanisms maintain inversion polymorphisms (Dobzhanky 1950, Dobzhanky 1970, Kirkpatrick and Barton 2006). According with Dobzhanky the evolution and maintenance of inversions is due to their role in maintaining co-adapted gene complexes - sets of adaptive genes that interact epistatically. According to his hypothesis, inversions are selected because in heterokaryotypes they prevent recombination that would undesirably break co-adapted combinations of alleles. More recently Kirkpatrick and Barton (2006) presented an alternative explanation for the selective advantage of inversions, also involving their role in reducing recombination between locally adapted genes, but where epistasis is not required. How to disentangle between these two hypotheses? While epistasis may occur in both models, its absence is against Dobzhansky's model. Another expectation of Dobzhansky model is that the genetic content of inversions differs between populations. Analysing these features may thus contribute to the debate (Schaeffer et al. 2003, Simões et al. 2012, Santos et al. 2016). I found indications that populations are genetically differentiated in chromosomes with the same inversion. This finding, together with the different set of SNPs with signs of selection between populations, having a high percentage associated with the inversion, suggests that the genetic content of inversions differs between populations. The genetic content of inversions seems also to be changing during laboratory adaptation. My data is in a way favorable of Dobzhanky but we still lack much information including a complete reference

genome with the localization of the breakpoints and more analysis on linkage between SNPs and the inversions. A recent study by Santos et al. (2016) involving populations of *D. subobscura* founded from Adraga a few years before this study, analyzed the evolutionary dynamics of a few microsatellites. They found some indication that epistatic selection was at play involving the genetic content of inversions. But very few markers were used in that study. Widening the study to many SNPs as we have with the RAD-seq data is a must, but we still lack much information on their location relative to inversions breakpoints, essential to deepen the study of the genomic evolution of inversions.

The analysis of chromosomes with A2 inversion allows to take more accurate conclusions about genomic changes across populations and generations, since we had more individuals sequenced in all samples having that inversion. Interestingly the total SNPs detected in chromosomes with A2 inversion were more differentiated between generations in Gro than between generations in Ad (Appendix 4 – first line A2, corresponding to global set of SNPs), that is, the chromosomes with A2 inversion changed more throughout time in Gro than in Ad. This is in accordance with Seabra et al. (2017), that found that the majority of the candidate SNPs of Gro were located in chromosome A, in contrast with Ad (with the majority of candidate SNPs in the O and E chromosome).

As expected and seen for the other arrangements, candidate SNPs defined for a given population did not change so much between generations in the other population (Appendix 4). Importantly, as mentioned above, the differentiation was also higher in the other population compared with the one using the all set of SNPs (Appendix 4). This finding suggests that though the SNPs under selection were not in common between populations, some of them do respond to selection in both.

Initial differentiation between populations was always higher for SNPs under selection than for the whole set of SNPs. Importantly that differentiation even increases across generations, again more for candidate SNPs (Appendix 5). This may be in part due to the fact that some SNPs are differentially fixed between populations and thus do not change across generations in one of them. Such lack of initial genetic variability contrasts with what was seen by Seabra et al. (2017) with poolseq, and may contribute to the lack of genetic convergence observed between Ad and Gro. That is, some SNPs that gave signs of selection in one population were fixed in the other and obviously did not respond due to that fact and not due to different genetic background (e.g. epistasis, see discussion in Seabra et al. 2017). More detailed analysis, e.g. by simulating drift, as well as analysing the actual evolutionary trajectories of candidate SNPs is required to deepen our understanding of what causes different adaptive genomic evolution in our populations. This may be in part due to the fact that some SNPs are differentially fixed between populations and thus don't change across generations in one of them. Such lack of initial genetic variability may contribute to the lack of genetic convergence observed between Ad and Gro, and contrasts with what was seen by Seabra et al. (2017) with poolseq.

The increase in the differentiation between populations across generations is higher for A2 inversion than for the other arrangements. Although this could be generated by the small sample size of other arrangements (O_{3+4} and O_{ST}, Appendix 5), the sex chromosome (chromosome A in *Drosophila subobscura*) has been reported as presenting more divergence than autossomes (Wong Miller et al. 2017) and faster rates of evolution (Musters 2006), consistent with studies of Charlesworth et al. (1987), Thornton and Long (2002), Torgerson and Singh (2003) and Richards et al. (2005). Wong Miller et al. (2017) suggested that the sex chromosome play an important role in population differentiation within species. However, they were not able to clarify the reason for overall patterns of increased sex chromosome-linked divergence in their study. The fast evolution of A chromosome increases the confidence that the non-convergence observed during the, 25 generations analysed, is not just a matter of time.

It was previously shown that, in both Ad and Gro there is an increase in the frequency of the A_2 inversion during laboratory adaptation, bigger in Ad though also present in Gro (Fragata et al. 2014a). This indicates that this inversion may have a higher adaptive value relative to the others, in particular A_{ST}. In this work, I observed an increase in the number of candidate SNPs associated with the inversion over time in both Ad and Gro. This suggests that selection increases the differences in the genetic content of the inversions, which is an interesting finding that deserves future analysis. Our present data suggests an underlying complex evolutionary dynamic, with differential selective pressures playing a role not only in the changes of inversion frequencies but also in the specific genetic content within the inverted region. In a sense this goes in accordance with the co-adapted complex hypothesis of Dobzhansky (Dobzhansky 1950, 1970) that involves non-linear, epistatic interactions between genes under selection.

4.2. Comparing the conclusions of the Pool-seq and the RAD-seq data

The analysis of the RAD-seq data indicates no convergence, either at the genome-wide level or for candidate SNPs. This finding is in accordance with what was reported in the pool-seq study of Seabra et al. (2017), that also observed no convergence for genome-wide and candidate SNPs for the same populations and generations. There were almost no common candidate SNPs between Ad and Gro in the RAD-seq study, again as well as in the pool-seq data. In both studies, partly the non-convergence, as well as the disparities in the sets of candidate SNPs between populations, may be due to the fact that the allele that is being detected in one population may be major in the other (and thus does not pass the filter of minor allele increasing). This could prevent convergence to happen. Nevertheless, in contrast with Seabra et al. (2017), in part the non-convergence observed here for Rad-Seq, particularly for A2 chromosomes, may be because candidate SNPs in one population were fixed in the other. Also, the allele that is being detected in one population may be major in the other. This could prevent convergence from happen.

I did not detect candidate SNPs in common between the two approaches. This could be due to the fact that: 1) in the Pool-seq study, sequences are for the whole genome, whereas for RAD-seq only regions that were cut by enzyme, so potentially fewer SNPs would be detected by RAD-seq; 2) in the Pool-seq analysis the total of SNPs were used to search for candidate SNPs whereas in the RAD-seq approach we analysed the SNPs present in each chromosome, and in these sets the specific CMH cut-off lines used were different; 3) the methods used to identify candidate SNPs under positive selection were different, particularly for the RAD-seq approach simulations were not done.

Although I did not detect candidate SNPs in common between the two approaches, I detected common scaffolds. Comparing the pool-seq set of candidate SNPs with the 6 candidate SNPs sets detected in the RAD-seq analysis (that is, for $O_{\underline{3+4}}$, O_{ST} and A_2, for both Ad and Gro) I obtained: chromosomes with $O_{\underline{3+4}}$ arrangement SNP set - 8 scaffolds in common with candidate SNPs in Ad and no one with candidate SNPs in Gro); chromosomes with O_{ST} inversion SNP set - no scaffolds in common with candidates in Ad but 2 scaffolds with candidates in Gro; chromosomes with A_2 inversion SNP set - 1 scaffold in Ad and 33 in Gro. More detailed analysis of these scaffolds, including linkage disequilibrium and how they differ across populations and generations will be among the priorities of future analysis (see below).

4.3. Future perspectives

Although this work gave further inside on the genomics of adaptive evolution of populations and the evolution of the genomic content of arrangements, there are many more analyses that need to be done to clarify several issues.

In the immediate future, RAD-seq data will be further explored, namely: 1) simulations of drift vs selection, to more accurately detect SNPs under selection; 2) calculate linkage disequilibrium involving candidate SNPs localized within A2 taking advantage of known haplotypes, e.g. evolution of linkage disequilibrium within A2 and how it differs between populations; 3) a detailed analysis of genomic changes on candidate SNPs of A2 that were not fixed in G6 of the other population and are associated with inversions – are they showing the same or different dynamics? 4) All the previous analysis of RAD-seq data will be applied to other inversions.

While *D. subo*bscura genome is not fully assembled, it will be important to develop or improve bioinformatics tools, namely: 1) Improve the database that I populated by inserting more available data (e.g. new published sequences, scaffold information from draft reference genome) and links to other available databases, as well as improving user interaction with the database; 2) Automatize the process of characterizing the type of mutations, that requires linking annotated DNA sequencing information of other species with that of *D. subobscura*;

When the full assembled reference genome and the mapping of inversions are available, an analysis of the patterns of variation and differentiation along the chromosomes and inside/outside inversions will allow a better understanding of the genetic mechanisms underlying the evolution of inversions and the genomics of adaptation in general.

To conclude, this work is innovative by using an 'evolve and resequencing' approach combining Pool-seq and RAD-seq analysis of the same populations given that the two studies are complementary. This powerfull combined approach allowed to tackle a most relevant issue, the role of history in genomic evolution of populations with contrasting biogeographical history. In particular it allowed for the first time to address genome-wide relevance of the evolution of inversions in *Drosophila subobscura*, a species highly polymorphic for inversions. With this master project, I developed new bioinformatic tools that were essential and will help further analysis in this and other species. An important realization is that a fundamental step in bioinformatic analysis is the definition of the set of parameters that will condition all further analysis and affect the conclusions. This is the most relevant issue because we are witnessing a boom in the number of genome-wide studies that deal with populations data, but that still does not reflect the required maturity on this subject and that will hopefully occur in future. It will be necessary careful analysis of the data, careful deal with inputs and file conversion. Summing up, this work allowed further insights on the genomics of adaptation giving rise to many new questions and avenues of research.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF. 2000. The Genome Sequence of *Drosophila melanogaster*. Genetics 287:2185–2195.

Ayala FJ, Serra L, Prevosti A. 1989. A grand experiment in evolution: the *Drosophila subobscura* colonization of the Americas. Genome 31:246–255.

Bailey SF, Bataillon T. 2016. Can the experimental evolution programme help us elucidate the genetic basis of adaptation in nature? Mol. Ecol. 25:203–218.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:1–7.

Balanyà J, Solé E, Oller JM, Sperlich D, Serra L. 2004. Long-term changes in the chromosomal inversion polymorphism of *Drosophila subobscura*. II. European populations. J. Zool. Syst. Evol. Res. 42:191–201.

Baldwin-Brown JG, Long AD, Thornton KR. 2014. The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. Mol. Biol. Evol. 31:1040–1055.

Barton NH, Keightley PD. 2002. Understanding Quantitative Genetic Variation. Nat. Rev. Genet. 3:11–21.

Becker RA, Chambers JM, Wilks AR. 1988. *The New S Language*. Wadsworth & Brooks/Cole.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29: 1165–1188.

Bergland AO, Tobler R, González J, Schmidt P, Petrov D. 2015. Secondary contact and local adaptation contribute to genome- wide patterns of clinal variation in *Drosophila melanogaster*. Mol. Ecol. 25:1157–1174.

Blount ZD, Borland CZ, Lenski RE. 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc. Natl. Acad. Sci. 105:7899–7906.

Buckling A, Craig Maclean R, Brockhurst MA, Colegrave N. 2009. The Beagle in a bottle. Nature 457:824–829.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: An analysis tool set for population genomics. Mol. Ecol. 22:3124–3140.

Charlesworth B, Coyne JA, Barton NB. 1987. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. Am. Nat. 130:113–146.

Cohan FM, Hoffmann AA. 1986. Genetic divergence under uniform selection. II. Different responses to selection for knockdown resistance to ethanol among *Drosophila melanogaster* populations and their replicate lines. Genetics 114:145–164.

Cohan FM, Hoffmann AA. 1989. Uniform Selection as a Diversifying Force in Evolution : Evidence from *Drosophila*. Am. Nat. 134:613–637.

Conte GL, Arnegard ME, Peichel CL, Schluter D. 2012. The probability of genetic parallelism and convergence in natural populations. Proc. R. Soc. B Biol. Sci. [Internet] 279:5039–5047.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Davey JL, Blaxter MW. 2011. RADseq: Next-generation population genetics. Brief. Funct. Genomics 9:416–423.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. [Internet] 12:499–510.

Dettman JR, Rodrigue N, Melnyk AH, Wong A, Bailey SF, Kassen R. 2012. Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. Mol. Ecol. 21:2058–2077.

de Visser JAGM, Krug J. 2014. Empirical fitness landscapes and the predictability of evolution. Nat Rev

Genet. 15:480–490.

Dobzhansky T, Epling C. 1948. The Suppression of Crossing Over in Inversion Heterozygotes of *Drosophila Pseudoobscura*. Proc. Natl. Acad. Sci. 34:137–141.

Dobzhansky T Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of *Drosophila pseudoobscura*. 1950. Genetics 35:288–302.

Dobzhansky T Genetics of the Evolutionary Process. (Columbia University Press, 1970).

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl. Acad. Sci. [Internet] 107:16196–16200.

Farlow A, Dolezal M, Hua L, Schlötterer C. 2012. The genomic signature of splicing-coupled selection differs between long and short introns. Mol. Biol. Evol. 29:21–24.

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

Fouet C, Gray E, Besansky NJ, Costantini C. 2012. Adaptation to aridity in the malaria mosquito anopheles gambiae: Chromosomal inversion polymorphism and body size influence resistance to desiccation. PLoS One 7.

Fragata I, Lopes-Cunha M, Bárbaro M, Kellen B, Lima M, Santos MA, Faria GS, Santos M, Matos M, Simões P. 2014a. How much can history constrain adaptive evolution? A real-time evolutionary approach of inversion polymorphisms in *Drosophila subobscura*. J. Evol. Biol. 27:2727–2738.

Fragata I, Simões P, Lopes-Cunha M, Lima M, Kellen B, Bárbaro M, Santos J, Rose MR, Santos M, Matos M. 2014b. Laboratory selection quickly erases historical differentiation. PLoS One 9.

Fuller ZL, Haynes GD, Richards S, Schaeffer SW. 2016. Genomics of natural populations: How differentially expressed genes shape the evolution of chromosomal inversions in *Drosophila pseudoobscura*. Genetics 204:287–301.

Gilchrist GW, Huey RB, Balanyà J, Pascual M, Serra L. 2004. A time series of evolution in action: a latitudinal cline in wing size in South American *Drosophila subobscura*. Evolution (N. Y). 58:768–780.

Gilmore WJ. 2006. Beginning PHP and MySQL 5. Second Edi. Apress

Gramates LS, Marygold SJ, Dos Santos G, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB, et al. 2017. FlyBase at 25: Looking to the future. Nucleic Acids Res. 45:D663–D671.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. [Internet] 41:95–98.

Hoffmann AA, Rieseberg LH. 2008. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? Annu. Rev. Ecol. Evol. Syst. 39:21–42.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6.

Huey RB, Gilchrist GW, Carlson ML, Berrigan D, Serra L. 2000. Rapid Evolution of a Geographic Cline in Size in an Introduced Fly. Science (80-.). [Internet] 287:308–309.

Ito K, Katsuma S, Kuwazaki S, Jouraku A, Fujimoto T, Sahara K, Yasukochi Y, Yamamoto K, Tabunoki H, Yokoyama T, et al. 2016. Mapping and recombination analysis of two moth colour mutations, Black moth and Wild wing spot, in the silkworm Bombyx mori. Heredity (Edinb). 116:52–59.

Jacob F. 1977. Evolution and Tinkering. Science (80-.). 197:1161–1166.

Kamdem C, Fouet C, White BJ. 2017. Chromosome arm specific patterns of polymorphism associated with chromosomal inversions in the major African malaria vector, Anopheles funestus. Mol. Ecol.:1–15.

Kapun M, Fabian DK, Goudet J, Flatt T. 2016. Genomic Evidence for Adaptive Inversion Clines in *Drosophila melanogaster*. Mol. Biol. Evol. 33:1317–1336.

Kirkpatrick M, Barton N. 2006. Chromosome inversions, local adaptation and speciation. Genetics 173: 419–434.

Koske TH, Maynard Smith J. 1954. Genetics and cytology of *Drosophila subobscura*. X. The fifth linkage group. Journal of Genetics 52:521–541.

Krimbas C. 1992. The inversion polymorphism of *Drosophila subobscura*. In: *Drosophila* Inversion Polymorphism (C. Krimbas & J. Powell, eds), pp. 127–220. CRC Press, Boca Raton, FL.

Krimbas CB. 1993. *Drosophila subobscura*. Biology, Genetics and Inversion Polymorphism. Verlag Dr Kovac: Hamburg.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. [Internet] 10:R25.

Lobkovsky AE, Koonin E V. 2012. Replaying the tape of life: Quantification of the predictability of evolution. Front. Genet. 3:1–8.

Long A, Liti G, Luptak A, Tenaillon O. 2015. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nat. Rev. Genet. 513:569–573.

Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guillier F, Janssens H, Ji W, Mclaren P, North P, et al. 2007. FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. Genome Biol. 8:R129.

Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 22: 719–748.

Menozzi P, Krimbas CB. 1992. The inversion polymorphism of *D. subobscura* revisited: Synthetic maps of gene arrangement frequencies and their interpretation. J. Evol. Biol. 5: 625-641.

Merilä J, Hendry AP. 2014. Climate change, adaptation, and phenotypic plasticity: The problem and the evidence. Evol. Appl. 7:1–14.

Munté A, Rozas J, Aguadé M, Segarra C. 2005. Chromosomal inversion polymorphism leads to extensive genetic structure: a multilocus survey in *Drosophila subobscura*. Genetics 169: 1573–1581.

Musters H, Huntley MA, Singh RS. 2006. A genomic comparison of faster-sex, faster-X, and faster-male evolution between *Drosophila melanogaster* and *Drosophila pseudoobscura*. J. Mol. Evol. 62:693–700.

Orgogozo V. 2015. Replaying the tape of life in the twenty-first century. Interface Focus 5:20150057.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *drosophila*. Mol. Biol. Evol. 27:1226–1234.

Pegueroles C, Aquadro CF, Mestres F, Pascual M. 2013. Gene flow and gene flux shape evolutionary patterns of variation in *Drosophila subobscura*. Heredity (Edinb). 110:520–529.

Pegueroles G, Mestres F, Argemí M, Serra L. 1999. Phenotypic plasticity in colonizing populations of *Drosophila subobscura*. Genet. Mol. Biol. 22:511–516.

Plucain J, Suau A, Cruveiller S, Médigue C, Schneider D, Le Gac M. 2016. Contrasting effects of historical contingency on phenotypic and genomic trajectories during a two-step evolution experiment with bacteria. BMC Evol. Biol. 16:86.

Prevosti A, Ribo G, Serra L, Aguade M, Balana J, Monclus M, Mestres F. 1988. Colonization of America by *Drosophila subobscura*: Experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism. Proc. Natl. Acad. Sci. 85:5597–5600

Rego C, Balanyà J, Fragata I, Matos M, Rezende EL, Santos M. 2010. Clinal patterns of chromosomal inversion polymorphisms in *Drosophila subobscura* are partly associated with thermal preferences and heat stress resistance. Evolution (N. Y). 64:385–397.

Rezende EL, Balanyà J, Rodríguez-Trelles F, Rego C, Fragata I, Matos M, Serra L, Santos M. 2010. Climate change and chromosomal inversions in *Drosophila subobscura*. Clim. Res. 43:103–114.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, et al. 2005. Comparative genome sequencing of. Genome 15:1–18.

Sambrook J, Russell DW (2001) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Santos J, Pascual M, Fragata I, Simões P, Santos MA, Lima M, Marques A, Lopes-Cunha M, Kellen B, Balanyà J, et al. 2016. Tracking changes in chromosomal arrangements and their genetic content during adaptation. J. Evol. Biol. 29:1151–1167.

Santos M. 2009. Recombination load in a chromosomal inversion polymorphism of *Drosophila subobscura*. Genetics 181:803–809.

Santos M, Céspedes W, Balanyà J, Trotta V, Calboli FCF, Fontdevila A, Serra L. 2005. Temperature-related genetic changes in laboratory populations of *Drosophila subobscura*: evidence against simple climatic-based explanations for latitudinal clines. Am. Nat. 165:258–273.

Schaeffer SW, Goetting-Minesky MP, Kovacevic M, Peoples JR, Graybill JL, Miller JM, Kim K, Nelson JG, Anderson WW. 2003. Evolutionary genomics of inversions in *Drosophila* pseudoobscura: Evidence for epistasis. Proc. Natl. Acad. Sci. 100:8319–8324.

Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU. 2015. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. Heredity (Edinb). 114:431–440.

Schlötterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. Nat. Rev. Genet. 15:749–763.

Seabra SG, Fragata I, Antunes MA, Faria GS, Santos MA, Sousa VC, Simões P, Matos M. 2017. Different genomic changes underlie adaptive evolution in populations of contrasting history. Mol. Biol. Evol. Online early, DOI:10.1093/molbev/msx247

Seeb LW, Waples RK, Limborg MT, Warheit KI, Pascal CE, Seeb JE. 2014. Parallel signatures of selection in temporally isolated lineages of pink salmon. Mol. Ecol. 23:2473–2485.

Simões P, Calabria G, Picão-Osório J, Balanyà J, Pascual M. 2012. The Genetic Content of Chromosomal Inversions across a Wide Latitudinal Gradient. PLoS One 7.

Simões P, Fragata I, Seabra SG, Faria GS, Santos MA, Rose MR, Santos M, Matos M. 2017. Predictable phenotypic, but not karyotypic, evolution of populations with contrasting initial history. Sci. Rep. 7:913.

Sperlich D, Feuerbach-Mravlag H, Lange P, Michaelidis A, Pentzos-Daponte A. 1977. Genetic load and viability distribution in central and marginal populations of *Drosophila subobscura*. Genetics 86: 835–848.

Spor A, Kvitek DJ, Nidelet T, Martin J, Legrand J, Dillmann C, Bourgais A, Vienne D De, Sherlock G, Sicard D. 2014. Phenotypic and genotypic convergences are influenced by historical contingency and environment in yeast. Evolution (N. Y). 68:772–790.

Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. 2012. The Molecular Diversity of Adaptive Convergence. Science (80-.). 335:457–461.

Teotónio H, Chelo IM, Bradic M, Rose MR, Long AD. 2009. Experimental evolution reveals natural selection on standing genetic variation. Nat. Genet. 41:251–257.

Teotónio H, Rose MR. 2000. Variation in the reversibility of evolution. Nature 408:463–466.

Thornton K, Long M. 2002. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. Mol. Biol. Evol. 19:918–925.

Tobler R, Franssen SU, Kofler R, Orozco-Terwengel P, Nolte V, Hermisson J, Schlötterer C. 2014. Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. Mol. Biol. Evol. 31:364–375.

Torgerson DG, Singh RS. 2003. Sex-linked mammalian sperm proteins evolve faster than autosomal ones. Mol. Biol. Evol. 20:1705–1709.

Travisano M, Mongold JA, Bennett AF, Lenski RE. 1995. Experimental tests of the roles of adaptation, chance, and history in evolution. Science (80-.). 267:87–90.

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

Vasemägi A. 2006. The Adaptive Hypothesis of Clinal Variation Revisited: Single-Locus Clines as a Result of Spatially Restricted Gene Flow. Genetics 173:2411–2414.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. Springer.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution (N. Y). 38:1358–1370.

Wong Miller KM, Bracewell RR, Eisen MB, Bachtrog D. 2017. Patterns of Genome-Wide Diversity and Population Structure in the *Drosophila athabasca* Species Complex. Mol. Biol. Evol. 34:1912–1923.

Wright S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc. 6th Int. Congress Genet. 1, 356–366.

Yin Z, Lan H, Tan G, Lu M, Vasilakos A V., Liu W. 2017. Computing Platforms for Big Biological Data Analytics: Perspectives and Challenges. Comput. Struct. Biotechnol. J. 15:403–411.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326–3328.

Appendix

Appendix 1 – List of all publications included in the *DsubSeqLoc* database.

Araúz PA, Peris-Bondia F, Latorre A, Serra L, Mestres F. 2011. Molecular evidence to suggest the origin of a colonization: *Drosophila subobscura* in America. Genetica 139:1477–1486.

Cirera S, Aguadé M. 1998. The sex-peptide gene (Acp70A) is duplicated in *Drosophila subobscura*. Gene 210:247–254.

Herrig DK, Modrick AJ, Brud E, Llopart A. 2014. Introgression in the *Drosophila subobscura-D. madeirensis* sister species: Evidence of gene flow in nuclear genes despite mitochondrial differentiation. Evolution (N. Y). 68:705–719.

Laayouni H, García-Franco F, Chávez-Sandoval BE, Trotta V, Beltran S, Corominas M, Santos M. 2007. Thermal evolution of gene expression profiles in *Drosophila subobscura*. BMC Evol. Biol. 7:42.

Moltó MD, Pascual L, Martínez-Sebastián MJ, De Frutos R. 1992. Genetic analysis of heat shock response in 3 *Drosophila* species of the obscura group. Genome 35:870–880.

Munté A, Aguadé M, Segarra C. 2000. Nucleotide Variation at the yellow Gene Region is not Reduced in *Drosophila subobscura*: A Study in Relation to Chromosomal Polymorphism. Names 17:1942–1955.

Munté A, Rozas J, Aguadé M, Segarra C. 2005. Chromosomal inversion polymorphism leads to extensive genetic structure: A multilocus survey in *Drosophila subobscura*. Genetics 169:1573–1581.

Navarro-Sabaté À, Aguadé M, Segarra C. 1999. The Relationship Between Allozyme and Chromosomal Polymorphism Inferred From Nucleotide Variation at the Acph-1 Gene Region of *Drosophila subobscura*. Genetics 153:871–889.

Orengo DJ, Puerma E, Papaceit M, Segarra C, Aguadé M. 2015. A molecular perspective on a complex polymorphic inversion system with cytological evidence of multiply reused breakpoints. Heredity (Edinb). 114:610–618.

Papaceit M, Segarra C, Aguadé M. 2013. Structure and population genetics of the breakpoints of a polymorphic inversion in *Drosophila subobscura*. Evolution (N. Y). 67:66–79.

Pegueroles C, Aquadro CF, Mestres F, Pascual M. 2013. Gene flow and gene flux shape evolutionary patterns of variation in *Drosophila subobscura*. Heredity (Edinb). 110:520–529.

Pegueroles C, Ferrés-Coy A, Martí-Solano M, Aquadro CF, Pascual M, Mestres F. 2016. Inversions and adaptation to the plant toxin ouabain shape DNA sequence variation within and between chromosomal inversions of *Drosophila subobscura*. Sci. Rep. 6:23754.

Penalva LOF, Sakamoto H, Navarro-Sabaté A, Sakashita E, Granadino B, Segarra C, Sánchez L. 1996. Regulation of the gene Sex-lethal: a comparative analysis of *Drosophila melanogaster* and *Drosophila subobscura*. Genetics 144:1653.

Pinsker W, Sperlich D. 1984. Cytogenetic mapping of enzyme loci on chromosomes J and U of *Drosophila subobscura*. Genetics 108:913–926.

Pratdesaba R, Segarra C, Aguadé M. 2015. Inferring the demographic history of *Drosophila subobscura* from nucleotide variation at regions not affected by chromosomal inversions. Mol. Ecol. 24:1729–1741.

Puerma E, Orengo DJ, Aguadé M. 2016. The origin of chromosomal inversions as a source of segmental duplications in the Sophophora subgenus of *Drosophila*. Sci. Rep. 6:30715; doi: 10.1038/srep30715.

Puerma E, Orengo DJ, Aguadé M, Sturtevant AH, Wesley CS, Eanes WF, Andolfatto P, Wall JD, Kreitman M, Cáceres M, et al. 2016. Multiple and diverse structural changes affect the breakpoint regions of polymorphic inversions across the *Drosophila* genus. Sci. Rep. 6:36248.

Puerma E, Orengo DJ, Salguero D, Papaceit M, Segarra C, Aguadé M. 2014. Characterization of the breakpoints of a polymorphic inversion complex detects strict and broad breakpoint reuse at the molecular level. Mol. Biol. Evol. 31:2331–2341.

Sánchez-Gracia A, Rozas J. 2011. Molecular population genetics of the OBP83 genomic region in *Drosophila subobscura* and D. guanche: contrasting the effects of natural selection and gene arrangement expansion in the patterns of nucleotide variation. Heredity (Edinb). 106:191–201.

Santos J, Serra L, Solé E, Pascual M. 2010. FISH mapping of microsatellite loci from *Drosophila subobscura* and its comparison to related species. Chromosom. Res. 18:213–226.

Appendix 2 –Number of individuals per replicate and generation with $O_{3+4,}$ O_{ST} and A_2 arrangements.

	O_{3+4}	O_{ST}	A_2
Ad1G6	19	1	30
Ad1G25	13	18	30
Ad2G6	14	2	31
Ad2G25	26	6	39
Ad3G6	16	2	35
Ad3G25	31	1	37
Gro1G6	14	20	10
Gro1G25	1	29	21
Gro2G6	12	21	8
Gro2G25	6	24	14
Gro3G6	4	28	9
Gro3G25	1	26	18
TOTAL	157	178	282

Appendix 3 – CMH cuttoff values comparison between A) Pool-seq and B) RAD-seq.

A) Pool-seq	SNPset	Ad G6G25	Gro G6G25
	-Log P	7.6	7.1

B) RAD-seq	SNPset	Ad O_{3+4} G6G25	Gro O_{3+4} G6G25	Ad O_{ST} G6G25	Gro O_{ST} G6G25	Ad A_2 G6G25	Gro A_2 G6G25
	-Log P	13.5	6.6	15.4	11.4	12.0	17.3

Appendix 4 – Average differentiation between generations of Ad and Gro in all SNPsets.

arrangement	SNPset	average Ad	average Gro
O_{3+4}	Global	0.044	0.000
	under selection Ad	0.220	0.000
	under selection Gro	0.000	0.159
O_{ST}	Global	0.000	0.020
	under selection Gro	0.036	0.158
A_2	Global	0.026	0.079
	under selection Ad	0.136	0.113
	under selection Gro	0.036	0.389

Appendix 5 – Differentiation between Ad and Gro populations at each generation (G6 and G25) and change of differentiation between generations; the latter calculation does not always correspond to the difference of F_{ST} presented at each generation: F_{ST}G6 and F_{ST} G25 were calculated with all replicate populations with information available for that generation, while the difference was estimated with replicate data available in both generations. Negative F_{ST} values were considered zero, as in general more negative values does not correspond to less differentiated populations.

arrangement	SNPset	no. of SNPs	F_{ST}G6	F_{ST}G25	F_{ST}G25 – F_{ST}G6
$O_{\underline{3+4}}$	Global	40210	0.028	0.087	0.047
	candidate SNPs in Ad	36	0.041	0.118	0.083
	candidate SNPs in Gro	2	0.107	0	-0.091
O_{ST}	Global	40210	0	0.067	0.047
	candidate SNPs in Gro	64	0	0.159	0.093
A_2	Global	24921	0.017	0.101	0.084
	candidate SNPs in Ad	56	0.054	0.159	0.105
	candidate SNPs in Gro	100	0.062	0.534	0.472