Universitat Autònoma de Barcelona

Population genomics in Drosophila melanogaster : a bioinformatics approach

Sergi Hervás Fernández

DOCTORAL THESIS

Population genomics in *Drosophila melanogaster*: a bioinformatics approach

Author

Sergi Hervás Fernández

Supervisors

Dr. Antonio Barbadilla Prados

Dr. Sònia Casillas Viladerrams

Departament de Genètica i de Microbiologia
Facultat de Biociències
Universitat Autònoma de Barcelona

2018

Population genomics in *Drosophila melanogaster*: a bioinformatics approach

Genómica de poblaciones en *Drosophila melanogaster*: una aproximación bioinformática

Genòmica de poblacions a *Drosophila melanogaster*: una aproximació bioinformàtica

Memòria presentada per Sergi Hervás Fernández
per a optar al grau de Doctor en Genètica per
la Universitat Autònoma de Barcelona

Sergi Hervás Fernández	Dr. Antonio Barbadilla Prados	Dra. Sònia Casillas Viladerrams
Autor	***Director i Tutor***	***Directora***

Bellaterra, 19 de Setembre de 2018

Abstract

High-throughput sequencing technologies are allowing the description of genome-wide variation patterns for an ever-growing number of organisms. However, we still lack a thorough comprehension of the relative amount of different types of genetic variation, their phenotypic effects, and the detection and quantification of distinct selection regimes acting on genomes. The recent compilation of more than one thousand of worldwide wild-derived *Drosophila melanogaster* genome sequences reassembled using a standardized pipeline (*Drosophila Genome Nexus*, DGN, Lack et al. 2015, 2016) provides a unique resource to test molecular population genetics hypotheses, and ultimately understand the evolutionary dynamics of genetic variation in the populations. Besides, the increasing amount of genomic data available requires the continuous development and optimization of bioinformatics tools able to handle and analyze such information. Thus, the development and implementation of new biologically-oriented software addressing several steps from data acquisition, filtering, processing, display or analysis to the final reporting step is a constantly growing need, especially in fields dealing with large data sets, such as population genomics.

This thesis is conceived as a comprehensive bioinformatics and population genomics project. It is centered in the development and application of bioinformatics tools for the analysis and visualization of nucleotide variation patterns and the detection of selective events in the genome of *D. melanogaster*, using the DGN data. The main goal is accomplished in three sequential steps: (i) capture the evolutionary properties of the analyzed sequences (i.e., create a catalog of population genetics metrics) and implement a tool for the graphical display of such information; (ii) develop a statistical package for the computation of the diverse selection regimes acting on genomes (positive and purifying selection), and finally (iii) perform an initial population genomics analysis in *D. melanogaster* using the previously developed tools. The common approach applied to process the data, starting at the assembly of genome sequences and ending up at the estimates of population genetics metrics, allows performing, for the first time, a comprehensive comparison and interpretation of results using samples from five continents. Overall, this work provides a global overview of the nucleotide variation and adaptation patterns along the genome, and a general assessment of the relative impact of the major genomic determinants of genetic variation, in *Drosophila* meta-populations with different geographical origin.

Table of Contents

1. Introduction

1. Introduction

1.1 Biological evolution and molecular population genetics

The publication of Charles Darwin's work *The Origin of Species* (1859) supposed a revolutionary shifting in the way we understand and explain biological diversity. In his book, Darwin first introduced the concept of biological evolution, conceived as a population process where variation among individuals within a population is converted, through its magnification in time and space, in new populations, new species and, by extension, all biological diversity present, or that ever existed, on Earth (Lewontin 1974). Therefore, biological evolution is just the result of the process of change in populations through generations. There are two requirements for evolution to occur: (i) there must be variation in the phenotypic traits between individuals within a population and, (ii) this variation must be inheritable (at least partially) among generations (Lewontin 1970, Endler 1986).

In other words, evolution can be defined as a process of statistical transformation of groups of interbreeding individuals that share a common genetic pool which defines their phenotypes (i.e., Mendelian populations, the units of evolution). Thus, populations are defined by the distribution of alleles and genotypes from all their individuals.

Phenotypic traits are determined by a combination of the underlying genotypes and environmental pressures. DNA is the molecule that carries the genetic (genotypic) information (Avery et al. 1944). One of its most important properties is that it is intrinsically mutable, meaning that it has the potential to originate new genetic variants which can be accurately replicated and transmitted from generation to generation, providing the substrate on which evolution can occur.

These new genetic variants can contribute differentially to the survival or reproductive success of individuals within the populations (fitness differences). If they do so, then natural selection takes place.

Natural selection can be defined as a key process shaping the distribution of genetic variants with fitness effects in any population. Natural selection acts primary on the phenotype, the characteristics of the organism which actually interact with the environment, but the genetic (heritable) basis of any phenotype that gives that phenotype a reproductive advantage tend to become more common in a population. Over time, this process can result in populations that specialize for particular ecological niches (micro-evolution) and may eventually result in speciation (the emergence of new species, macro-evolution). Thus, natural selection only accounts for a subset of the evolutionary processes in which genetic variants differ in their fitness effects, but it is neither a necessary nor a sufficient condition for evolution to take place.

Population genetics provides the theoretical framework for explaining biological evolution from the variational paradigm. In words of Dobzhansky (1937): "The main aim of population genetics is the description and interpretation of genetic variation within and among populations". Hence, over the last century, population geneticists have developed an extensive theoretical framework which describes the dynamics of genetic variation (alleles and genotypes) in Mendelian populations and aims to ultimately understand the main determinants of the evolutionary rate.

The first population genetics mathematical model was proposed by G. H. Hardy and W. R. Weinberg in 1908 (the Hardy-Weinberg principle, Hardy 1908). This can be defined as the zero-force state model and served as a null model to explain the maintenance of genetic variation in populations. The principle states that in an ideal population and in the absence of any other evolutionary forces, allele frequencies would remain unchanged generation after generation once they reach the equilibrium state. The Hardy-Weinberg principle has the seven following assumptions underlying H-W equilibrium: (i) organisms are diploid, (ii) there is only sexual reproduction, (iii) generations are non-overlapping, (iv) mating is random, (v) population size is infinitely large, (vi) allele frequencies do not differ between sexes and, (vii) there is not any external force (gene flux, mutation or selection) affecting the population dynamics of new mutations.

Then, during the 30's-50's period, the population genetics theoretical principles and the dynamics and fundamental forces of evolution were established (Fisher 1930, Wright 1931, Haldane 1932, Kimura 1955). Hence, population genetics is conceived as a theory on which diverse forces can affect the allele frequencies in a population: mutation, migration, natural selection, recombination and genetic drift.

It was not until the late 60' that the first measures of genetic variation in the species *Drosophila pseudoobscura* (Lewontin & Hubby 1966) and humans (Harris 1966) were provided, which finally initiated the necessary dialog between data and theory. Since then, advances in molecular evolutionary genetics have subsequently enriched the field with many new concepts, terms, processes, molecular techniques, and statistical and computational methods (Casillas & Barbadilla 2017). However, the fundamental evolutionary forces established more than 50 years ago are still the essential explanatory factors used for understanding the population genetic basis of evolutionary change (Lynch & Walsh 2007, Charlesworth 2010).

1.1.1 Mutation as the ultimate source of genetic variation

Genetic variation is the cornerstone of the evolutionary process. This genetic variation is ultimately produced by mutations in the DNA molecule of single individuals, which can be replicated and inherited from generation to generation, producing the genetic and phenotypic diversity observed in nature.

Mutations are random or undirected events. This means they occur independently of whether they help or harm the individual in the environment in which it lives. Most mutations are lost in the first generations, but occasionally, some of them may increase in frequency through generations. This can be due to natural selection processes (because mutations are associated with higher fitness in the individuals that carry them) or just by random genetic drift.

Then, new mutations have three possible fates: (i) most are lost in a very short period of time, (ii) some of them keep segregating as polymorphisms (this is a genomic site which shows at least two different alleles in a population) and contribute to intra-population genetic variability, and finally, (iii) some of them reach fixation (one of the segregating alleles increases in frequency and replaces all the other alleles for that locus in the population).

For many years during the 19th century and the start of the 20th century, variation could only be studied for phenotypic traits due to the lack of proper techniques to analyze direct genetic changes. In general, the observed phenotypes are the final result of many interactions among genotypes (which are heritable) and the specific environment, but few cases are determined by discrete Mendelian variants. Hence, it is not possible to discern the single effect of each quantitative trait in the resulting phenotype, and few conclusions can be obtained in such analyses.

However, the advances in sequencing technologies during the last century provided ways to investigate the presence and effect of single genetic variants, allowing to test, using empirical data, the bulk of theoretical principles previously developed.

First studies of genetic variation were performed using gel electrophoresis of proteins (Lewontin & Hubby 1966, Harris 1966), but it was not until the milestone of sequencing technologies (Sanger & Coulson 1975, Maxam & Gilbert 1977) that genetic variation was estimated at the ultimate DNA sequence level (Kreitman 1983). Then, the parallelization of the Sanger method provided the field with dozens of sequenced genes in several species (Powell 1997). Nowadays, high throughput second or next-generation sequencing (NGS) techniques are providing big data samples of complete genome sequences of many individuals from natural populations of many species, allowing the development of catalogs of nearly all polymorphic variants in certain model species. For instance, the current human single nucleotide polymorphism (SNP) database lists 113,862,023 validated SNP (dbSNP, March 2018; `https://www.ncbi.nlm.nih.gov/SNP/`) and in *D. melanogaster*, >6,000,000 natural variants (SNPs and indels) have been described to date (Huang et al. 2014, Lack et al. 2016). In the coming years, population genomic data will continue to grow in both amounts of sequences and number of species (Ellegren 2014, Tyler-Smith et al. 2015), ensuring a constant progress of the field and allowing the detection of more and rarer genetic variants.

Several types of alterations in the genetic material (mutations) contribute to genetic diversity, from single nucleotide substitutions (SNPs) to duplications of the whole genome. Table 1.1 lists the most studied types of mutations so far.

Table 1.1: Common types of DNA mutations. Brief description of the most common types of alterations in the DNA classified in six major categories. (Adapted from Casillas 2008)

Type of mutation	Description
1. Single-nucleotide polymorphisms (SNPs)	Base substitution involving only a single nucleotide. Can be transitions or transversions. Coding-related mutations can be missense, nonsense, silent or splice-site mutations.
a) Transition	Substitution of one purine (A or G) by another, or one pyrimidine (C or T) by another.
b) Transversion	Substitution of a purine by a pyrimidine or vice-versa.
a) Missense	The new nucleotide alters the codons and produces an altered amino acid in the protein. Also called non-synonymous.
b) Nonsense	The new nucleotide changes a codon that specified an amino acid to one that stops prematurely the transcription, generating a truncated protein.
c) Silent	Replacement of one nucleotide by another which does not change the amino acid. Also called synonymous.
d) Splice-site	Mutations that alter the splice-site signals so that the intron cannot be removed from the RNA molecule, what results in an altered protein product.
2. Insertions and deletions (indels)	Extra base pairs that may be added (insertions) or removed (deletions) from the DNA. Many large indels result from the activity of transposable elements (TEs).
3. Variable number of tandem repeats (VNTR)	A locus that contains a variable number of short (2-8 nt for microsatellites, 7-100 nt for minisatellites) tandemly repeated DNA sequences that vary in length and are highly polymorphic. Microsatellites are also called short sequence repeats (SSRs) or short tandem repeats (STRs).
4. Copy number variation (CNV)	A structural genomic variant that results in confined copy number changes of DNA segments $\geq$ 1 kb (i.e., large duplications). They are usually generated by unequal crossing over between similar sequences.
5. Inversions	Change in the orientation of a piece of the chromosome, without any gain or loss of genetic material.
6. Translocations	Transfer of a piece of a chromosome to a non-homologous chromosome. Can often be reciprocal.

1.1.2 The (nearly) neutral theory of molecular evolution

It was not until the time when the genetic diversity of populations was beginning to be assessed by electrophoretic methods that the diverse theories regarding the major forces determining patterns of variation and the adaptive value of new mutations started to be tested using empirical data.

The large amount of genetic variation uncovered in nature, together with the previous observation that genetic differences accumulate linearly with time (Zuckerkandl & Pauling 1965) revealed that none of the formerly proposed hypothesis regarding the maintenance of variability in populations was able to properly explain these observations. Empirical studies showed that the rates of genetic variability would either impose a segregating load too great to be explained by balancing selection, as initially proposed by the balance hypothesis (Dobzhansky 1970, Ford 1971); or an insurmountable substitutional load in the case that directional positive selection was the major force driving amino acid substitutions (Casillas & Barbadilla 2017).

Motoo Kimura suggested a radical alternative explanation to account for the patterns of protein variation and substitution, known as the neutral theory of molecular evolution (Kimura 1968; Figure 1.1A). This theory postulates that the bulk of segregating polymorphic sites and fixed differences between species are selectively neutral (this is, they have a fitness effect of 0 and hence, are not affected by natural selection). The principal assertions and implications of Kimura's neutral theory of molecular evolution are (Kimura 1980, 1983) are:

1. Deleterious mutations are rapidly removed from the population, and adaptive mutations are rapidly fixed; therefore, most of the variation observed within species is the result of neutral mutations.

2. A steady-state rate at which neutral mutations are fixed in a population (k) equals the neutral mutation rate: $k = \mu_0$, where μ_0 is the neutral mutation rate, $\mu_0 = f_{neutral}\mu$, where $f_{neutral}$ is the proportion of all mutations that are neutral and

μ the intrinsic mutation rate by generation. If all mutations are neutral, then $\mu_0 = \mu$, and the average time between consecutive neutral substitutions is independent of population size ($1/\mu$).

3. The level of polymorphism in a population (θ) is a function of the neutral mutation rate and the effective population size (N_e): $\theta = 4N_e\mu$.

4. Polymorphisms are transient (on their way to a rapid loss or fixation) rather than balanced by selection. Larger populations (higher N_e) are expected to have a higher heterozygosity, as reflected in the greater number of alleles segregating at any time.

Under this model, the frequency dynamics of neutral variants in the population is determined by the rate of mutation and random genetic drift. Genetic drift is defined as the random sampling of gametes at each generation in a finite population, which results in a random fluctuation of allele frequencies across generations and ultimately, in the loss of genetic variation (Kimura 1968).

Kimura's neutral theory states that rates of protein evolution are proportional to generation time. However, empirical observations showed that they are proportional to absolute time in years instead of generations.

In this regard, Tomoko Ohta redefined Kimura's neutral theory by introducing the concept of nearly neutral mutations (Ohta 1973). Nearly neutral mutations are those which have slightly deleterious or advantageous fitness effects (a selection coefficient, s, around 0). This new class of mutations account for a major fraction of the observed polymorphisms.

Ohta's nearly neutral theory (1973) predicts that nearly neutral mutations are mostly eliminated by natural selection in large populations (those with a large effective population size, N_e), but that a substantial fraction of them behave as effectively neutral and are randomly fixed in small populations. Moreover, population size is generally inversely proportional to generation time. Because of this, the strength of purifying selection acting on slightly deleterious mutations and the generation time effect compensate and thus, protein evolution under Ohta's model is fairly insensitive to generation time.

In the 1990s, Ohta improved once again the model by including both slightly deleterious and slightly beneficial mutations (Ohta & Gillespie 1996; Figure 1.1B). This model has the following implications:

1. Mutations with fitness effects much smaller than $1/N_e$ are considered effectively neutral and their fate is directed mainly by genetic drift.

2. Mutations that have fitness effects on the order of $1/N_e$ are nearly neutral (slightly deleterious if s is negative, or slightly advantageous if s is positive), they have small effects on fitness, and their fate depends on a combination of natural selection and genetic drift.

3. Mutations with fitness effects $> 1/10N_e$ are strongly deleterious (if s is negative) or strongly advantageous (if s is positive), and their fates are mainly determined by natural selection.

The selection coefficient (s) is theoretically measured in the heterozygous state with the wild type, in the case of a diploid, randomly mating population. In a small population, the range between $-1/N_e$ and $1/N_e$ is larger than in a large population, and therefore there are more effectively neutral mutations. In contrast, in a large population most mutations are subject to some sort of natural selection. N_e is thus the key parameter determining the relative importance of selection vs. genetic drift. The range $|N_e s| = 1$ delimitates the decisive borderline: if $N_e s < 1$, genetic drift dominates, whereas if $N_e s > 10$, selection is the force which determines the fate of new mutations.

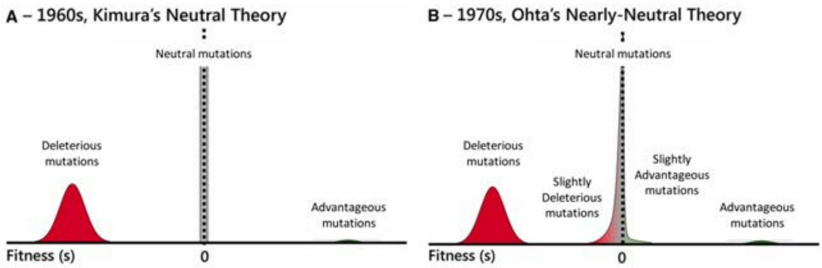


Figure 1.1: The (nearly) neutral theory of molecular evolution. Graphical representation of the distribution of fitness effects (DFE) for mutations according to the (nearly) neutral theory of molecular evolution. The fitness effect of new mutation is defined with the selection coefficient (s). At $s = 0$ the allele is said to be selectively neutral; as s increases so does its advantageous potential; in the same way, as s decreases so does the negative effect of a mutation. Different selection coefficients of mutations are colored in a gradient from maroon (strongly deleterious), red (slightly deleterious), gray (neutral), light green (slightly advantageous), and dark green (advantageous). **(A)** Kimura's neutral theory. Under this proposition, mutations are considered to be only neutral, advantageous or deleterious; **(B)** Ohta's nearly neutral theory. In this case, apart from neutral, advantageous and deleterious classes, some mutations are not completely neutral but either slightly advantageous or slightly deleterious.

1.1.3 Population dynamics of new mutations and the distribution of fitness effects

According to Otha's nearly neutral theory of molecular evolution (Ohta 1973, Ohta & Gillespie 1996), mutations are classified based on their individual fitness effects in five categories. However, classifying mutations into these discrete groups may not reflect properly the real fitness effects of mutations. In fact, there is a continuous distribution of selective effects, the distribution of fitness effects (DFE) (Eyre-Walker & Keightley 2007, Lanfear et al. 2014), such that the effects of mutations range from lethal or very deleterious, through slightly deleterious, neutral, slightly advantageous and strongly advantageous (Piganeau & Eyre-Walker 2003, Keightley & Eyre-Walker 2010).

Ultimately, the substitution rate of new mutations in populations is determined by both the distribution of fitness effects (DFE) of these new mutations and the population dynamics of each mutation from the moment it appears through time (this is, the changes in allele frequencies and their probability of fixation). This results in a mutation-selection balance which determines the fates of new variants (see above).

The DFE can be mathematically modeled as a function of the fitness (measured by the coefficient of selection, s) of new mutations entering in the population: $f(s)$ (Figure 1.2A). The probability of fixation of each mutation depends on the rate at which this class of mutations appear (μ per site per generation), the effective population size (N_e) and its fitness (s), and can be expressed as: $2N_e\ \mu(N_e, s)$ (Figure 1.2B).

Therefore, the rate of molecular evolution (K), which refers to the fixation rate averaged over all mutations entering a population, can be inferred with the combination of the DFE and population dynamics of new mutations (Figure 1.2). K informs about the rhythm at which species diverge along evolutionary time.

In a diploid population, mutations appear at a rate $2N_e\mu$ (Figure 1.2C). Each new mutation has a given selection coefficient (s) that is determined by its fitness effect on the individual (Figure 1.2A), and all mutations with this s, $f(s)$, appearing in a population of size N_e, have a probability of fixation $u(N_e, s)$ (Kimura 1957; Figure 1.2B). The selection coefficient (s) potentially ranges from $-\infty$ to $+\infty$. Thus, the overall molecular evolutionary rate (K) is determined by the general expression (Figure 1.2C):

$$K = 2N_e\mu \int_{-\infty}^{+\infty} u(N_e, s)\ f(s)\ ds$$

Hence, genetic variants can be generally divided into two classes: polymorphisms (the state in which multiple alleles exist for a same locus within the population) and fixations (a single allele for the same locus is shared by all the individuals within a population). Divergence is defined as the accumulation of fixed differences (alleles) among populations when two populations of the same species which are reproductively isolated derive in two different species.

Polymorphism and divergence rates inform about different but complementary stories about the past and present events of a population. Polymorphism captures the variation dynamics of the population at the precise moment in time it is observed, and only allows inferring events that have happened recently. Besides, divergence shows the fixations between species. This process requires longer periods of time and thus, provides information about more ancient events. The combined analysis of polymorphism and divergence rates is then one of the most powerful approaches to understand the influence of different population genetics forces modeling the patterns of molecular evolutionary change.

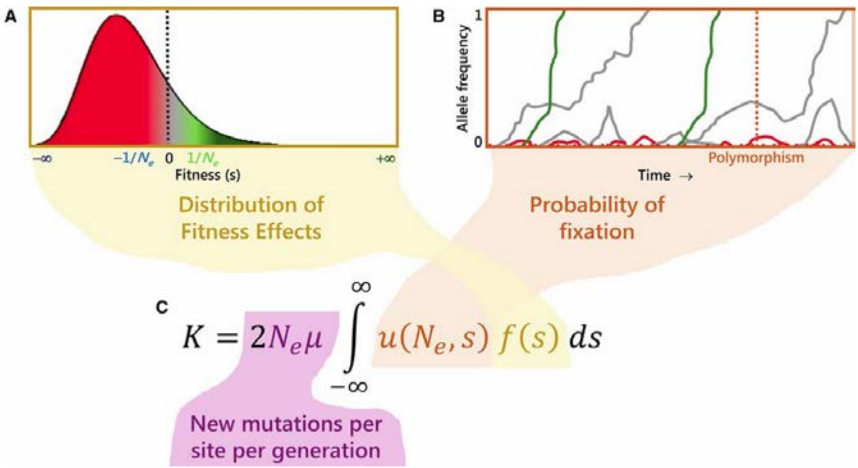


Figure 1.2: Population dynamics and the molecular evolutionary rate. The molecular evolutionary rate (K) can be expressed as a function of **(A)** the distribution of fitness effects (DFE), **(B)** the probability of fixation of new mutations entering the population (new variants that appear in a population start segregating and over time they can become fixed, frequency = 1, or disappear from the population, frequency = 0), and **(C)** the rate at which new mutations enter the population per site per generation. Different selection coefficients of mutations are colored in a gradient from maroon (strongly deleterious), red (slightly deleterious), gray (neutral), light green (slightly advantageous), and dark green (advantageous). (Retrieved from Casillas & Barbadilla 2017)

1.2 Detecting the genomic footprint of natural selection

1.2.1 Determinants of patterns of genetic variation

Recombination and linked selection

Recombination is a fundamental biological process in population genetics. The crossing-over of homologous chromosomes during meiosis results in the exchange of genetic material and the formation of new haplotypes, which provides offspring with new allelic combinations. Recombination rates are known to vary in magnitude and distribution among species (Ortiz-Barrientos et al. 2006, Auton et al. 2012), between populations within species (Consortium et al. 2015, Kong et al. 2010), between individuals within the same population (Kong et al. 2010, Coop et al. 2008) and even in different regions of the genome at different scales (Chan et al. 2012).

Therefore, the estimation of accurate rates of recombination along the genome is needed to understand the molecular and evolutionary mechanisms underlying recombination variation, allowing to assess the relationships between recombination and population genetics parameters to infer its relevance on genome evolution.

In this regard, diverse high-resolution recombination estimates have been proposed in *Drosophila melanogaster*, the model species for population genetics analyses (*see 1.3. Drosophila as a model organism for population genetics*). The first high resolution recombination map was provided by Fiston-Lavier et al. (2010), and it is based on comparing the genetic and physical maps to infer recombination rates along the genome. Later, two more high-resolution maps in this species were published:

- A statistical approach that infers the historical population recombination parameter ($\rho = 4N_e r$), from linkage disequilibrium patterns at multiple sites across the genome (Hudson 1987). Chan et al. (2012) developed a computationally intensive method to estimate ρ from nucleotide variation and LD rates which accounts for *Drosophila* specific genome properties (high density of SNPs, many missing allele calls, background recombination and pervasive adaptive evolution along the genome).

- An experimental-based approach developed by Comeron et al. (2012), using a new technique that integrates the power of classical genetics with NGS. First, recombinant advanced intercross lines (RAIL) were generated from 8 crosses among 12 wild-derived lines. Then, RAIL females were individually crossed to *D. simulans*, and the *D. melanogaster* haploid genome of single hybrid progeny was inferred. Thus, the authors generated the first integrated high-resolution description of the recombination patterns of both intra-genomic and population variation, distinguishing between crossing over (CO) and gene conversion (GC) events.

According to the neutral theory of molecular evolution, nucleotide variability and recombination rates should not be correlated, since the recombinational environment would not affect the number nor the frequency of neutral polymorphisms unless there is a positive correlation between recombination and the mutation rate (Lercher & Hurst 2002).

However, empirical population genetics studies demonstrated that there is a positive correlation between the local recombination rate and the polymorphism level, at different types of sites and variants (Begun & Aquadro 1992, Mackay et al. 2012), but no correlation was found between recombination and the rates of divergence, at least in *Drosophila*. These observations allowed discarding the hypothesis of the mutagenic nature of recombination and highlighted that recombination itself, rather than any other factor, seems to be the main process modulating the nucleotide diversity patterns along the genome. Since neutral theory cannot explain these observations, a selectionist alternative was proposed: linked selection (Begun & Aquadro 1992).

When genetic variants along a chromosome are in linkage disequilibrium (LD) (Lewontin & Kojima 1960), they tend to segregate together as a block because recombination has not been able to reshuffle them. Hence, the selection unit is not anymore a single mutation but each LD genomic block. Linked selection could explain the correlation between recombination and genetic variability because recombination can break the association between a selective target and its linked neutral alleles (Begun & Aquadro 1992). In addition, Birky & Walsh (1988) demonstrated that divergence is not affected long-term by this reduction of polymorphism level, which explains the lack of correlation between recombination and neutral divergence.

In the last years, several evolutionary models of recurrent linked selection have been applied to the genomic data available. These models also predict a positive correlation between recombination and polymorphism for all variants, including SNPs (Berry et al. 1991, Begun & Aquadro 1992) and indels (Huang et al. 2014). In addition, it has been observed that diversity increases with recombination rate but decreases with the density of functional sites (Begun et al. 2007, Shapiro et al. 2007), pinpointing that recombination rate via recurrent linked selection is the most likely explanation for the observed patterns of nucleotide variation in *Drosophila* (Mackay et al. 2012).

The Hill-Robertson interference

Analyses of genetic variation reveal that in species with a high N_e such as *Drosophila*, much of the genome is under purifying selection, and thus of functional importance, and that a large fraction of coding and non-coding differences between species are adaptive (Sella et al. 2009). In addition, genomes are populated by large numbers of segregating sites undergoing weak deleterious selection (Casillas & Barbadilla 2017). Therefore, purifying selection is pervasive in the genomes of such species. Indeed, Andolfatto (2005) estimated that in *D. melanogaster* the fraction of deleterious newly arising mutations was $\sim$ 94% at amino acid sites, $\sim$ 81% in untranslated regions (UTRs), $\sim$ 56% in introns, and $\sim$ 61% in intergenic regions. Later, Mackay et al. (2012) showed that averaged over the entire genome, 39.6% of the segregating sites are strongly deleterious (d), 58.5% are neutral or nearly neutral (f), and 1.9% are weakly deleterious (b). Authors also estimated that for non-synonymous sites these fractions were: $d = 77.6\%$, $b = 3.8\%$, and $f = 18.6\%$. Finally, Castellano

(2016) estimated, for the same class of sites: $d = 81\%$, $b = 3\%$, and $f = 15\%$. Altogether, these results reinforce the importance of purifying selection in *Drosophila*. Besides, a considerable fraction of amino acid substitutions are adaptive, with values ranging from 25% (Mackay et al. 2012), to $40 - 50\%$ (Andolfatto 2005), or as high as $\sim 57\%$ (Messer & Petrov 2013).

The combination of the large number of selected variants and linked selection implies that at any time there are genetic variants in LD simultaneously selected in the genome. These variants interfere with each other, inducing a cost of linkage known as the Hill-Robertson interference (HRi, Hill & Robertson 1966). HRi can be thus defined as the evolutionary consequence of selection acting simultaneously among multiple co-segregating sites in finite populations.

The effects of Hill-Robertson interference are illustrated in Figure 1.3. When two or more independent adaptive mutations occur in separate haplotypes without recombination (Figure 1.3A top; with HRi), only one of them can be fixed in the population and thus mutations compete, lowering the average rate of adaptive fixation. However, if recombination is high enough (Figure 1.3A bottom; in the absence of HRi), the two haplotypes can recombine to generate a new haplotype carrying both adaptive variants, which can be fixed. If adaptive and deleterious variants coexist in the same genomic block without recombination (Figure 1.3B top), some deleterious variants are dragged to fixation by linked adaptive ones, while the fixation rate of adaptive mutations is decreased due to the reduced efficacy of selection caused by linked deleterious variants. However, if recombination is high (Figure 1.3B bottom), deleterious alleles can be removed from the population, resulting in the fixation of the diverse adaptive variants and the elimination of the negative ones by purifying selection.

If the cost of linkage is real, the number of selected variants undergoing HRi will increase as recombination rate decreases. Thus, regions with higher recombination rates will show higher rates of adaptive fixation. Thereby, under this model, we expect higher rates of adaptive fixation and lesser fixation rate of deleterious variants as recombination rate increases.

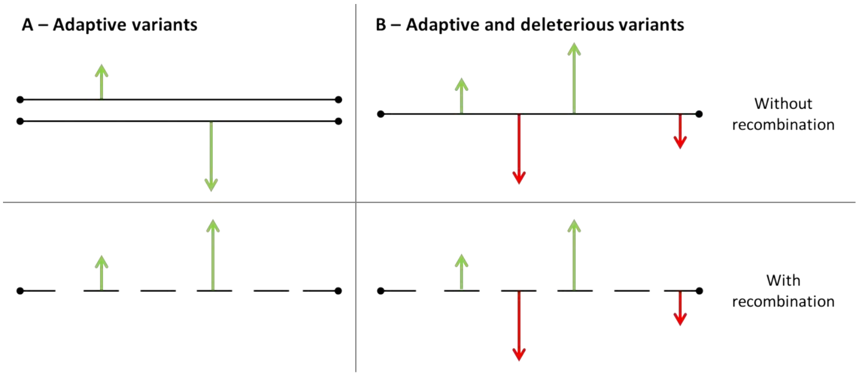


Figure 1.3: Representation of the HRi on selected sites in linkage disequilibrium. Arrows indicate adaptive (green) and deleterious (red) mutations, while their length indicates the intensity of selection. **(A)** Two or more adaptive mutations, segregating in separated low-recombining haplotypes competing for fixation (top), and with recombination, segregating independently and ultimately reaching both fixation (bottom). **(B)** Deleterious and adaptive variants coexisting in a low-recombinant haplotype (top) and with recombination, freely segregating towards fixation or lost (bottom). (Adapted from Barrón 2015).

In other words, the overall selection effectiveness is reduced (increased) as recombination decreases (increases). Therefore, HRi is predicted to be stronger in regions with lower recombination, a larger number of selected sites, and more intense selection (Comeron et al. 2008, Messer & Petrov 2013). In addition, the chromosome length affected by HRi depends on the recombination rate and the distribution of linked fitness variation along the chromosome.

Diverse population genomics studies performed in *Drosophila* confirm and reinforce the importance of HRi (Langley et al. 2012, Mackay et al. 2012, Campos et al. 2014, Castellano et al. 2016). Hence, if the HRi is common (Comeron et al. 2008), a central question is its magnitude. How much does HRi limit the molecular adaptation of a genome?

Castellano et al. (2016) specifically addressed this question by analyzing 6,141 autosomal protein-coding genes from the *Drosophila Genetic Reference Panel* (DGRP, Mackay et al. 2012) genome data. Results showed that the initially observed linear relationship between recombination and adaptation converged to an asymptotic threshold in recombination values around 2 cM/Mb, the same recombination threshold found by Mackay et al. (2012) when comparing recombination and nucleotide diversity rates.

This asymptote indicates that the cost of linkage disappears above a given recombination value and thus, mutations which appear in regions with recombination above this value can be assumed to segregate independently. This specific recombination threshold can be interpreted as the optimal recombination (r_{opt}) for the adaptation rate of a genome. Then, the determination of r_{opt} allows the quantification of the genome-wide impact of HRi on the adaptation rate, which correspond to the adaptation lost below the recombination threshold. Castellano et al. (2016) estimated that HRi reduces the evolutionary adaptation rate of the *D. melanogaster* genome by an average of 27%.

However, this study only focused on a North-American population of *D. melanogaster*, known to have experienced a recent bottleneck followed by a population expansion (Garud et al. 2015). Thus, the extent to which this proportion is representative of the whole species still remains unclear. The next logical step would be to quantify this effect in diverse wild-derived populations with different demographic and migratory histories in order to characterize which are the HRi load dynamics operating in *Drosophila* and their impact on the molecular evolutionary rate.

1.2.2 Genome-wide signatures of selection

Natural selection leaves signatures on the genome that can be now identified by taking advantage of the increasing amount of genomic information available.

During the process of fixation of an adaptive variant, closely linked alleles can also become more common by genetic hitchhiking (Smith & Haigh 1974), whether they are neutral or even slightly deleterious. Overall, this causes a reduction of the levels of genetic diversity in the region (Figure 1.4A). At the same time, new mutations appear and accumulate. These mutations have a low initial frequency, resulting in an excess of rare derived alleles in the selected area. In addition, if the fixation process of the adaptive variant is faster than the reduction of linkage disequilibrium between this selected mutation and linked variants by local recombination processes, the region will present long-range LD signals (Nielsen 2005, Franssen et al. 2015, Garud et al. 2015, Garud & Petrov 2016). In summary, selective sweeps leave signatures in the genome that include: (i) a reduction in the genetic diversity, (ii) a skew toward rare derived alleles and (iii) an increase in the LD.

Besides, a different selective process that also reduces the levels of genetic diversity in a genomic region is background selection (BGS) (Charlesworth et al. 1993, Braverman et al. 1995, Charlesworth et al. 1995). In this case, recurrent purifying selection eliminates chromosomes (haplotypes) carrying strongly deleterious mutations, producing a decrease genetic variation in that region due to the reduction of the number of chromosomes that contribute to the next generation (Figure 1.4B). In contrast to a hitchhiking event, it does not skew the distribution of rare polymorphisms nor generates long-range LD blocks. In this sense, the genomic footprint is identical to that of a reduction in population size but focused only on a specific linked genomic region (Charlesworth et al. 1993).

Finally, balancing selection and local adaptation leave other particular signatures of selection in the genome. These signatures include haplotypes at an intermediate frequency, with strong population differentiation; and a high level of LD with respect to variants at surrounding sites (Charlesworth et al. 1997).

1.2.3 Tests of selection

Since direct genetic variation data became available on a regular basis, one major issue in population genetics has been to detect and measure the amount of positive selection in the genome from polymorphism and/or divergence levels. Therefore, several tests have been devised to address this challenge, allowing to investigate the micro- and macro-evolutionary histories (i.e., within and between species) of a broad range of organisms (Vitti et al. 2013). Table 1.2 includes a brief description of some of the most used tests of neutrality developed and applied over the last years.

Table 1.2: Tests of selection. Most used tests of neutrality classified based on the type of selective events they are able to detect and the type of genomic data that they use.

	Based on	Test	Description	Reference
Micro-evolution	The allele frequency spectrum and /or levels of variability	Tajima's D	Number of nucleotide polymorphisms with the mean pairwise difference between sequences	Tajima (1989)
		Fu and Li's D	Number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants	Fu & Li (1993)
		Fu and Li's F	Number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences	Fu & Li (1993)
		Fay and Wu's H	Number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies	Fay & Wu (2000)
	Linkage disequilibrium	EHH	Extended haplotype homozygosity: measurement of the decay of LD between loci with distance	Sabeti et al. (2002)
		LRH	Long-range haplotype test, based on the frequency of alleles in regions of long-range LD	Sabeti et al. (2002)
		HHT	Hudson haplotype test: detection of derived and ancestral alleles on unusually long haplotypes	Hudson et al. (1994)
		iHS	Integrated haplotype score, based on the frequency of alleles in regions of high LD	Voight et al. (2006)
	Population differentiation	G_{ST}	Analysis of gene diversity (heterozygosity) within and between subpopulations	Nei (1973)
		F_{ST}	Average levels of gene flow based on allele frequencies, under the infinite-sites model	Hudson et al. (1992)
		$XP-CLR$	Cross-population composite likelihood ratio test, based on allele frequency differentiation across populations	Chen et al. (2010)
Macro-evolution	Comparisons of polymorphism and / or divergence between different classes of sites	D_N/D_S, K_A/K_S	Ratio of non-synonymous to synonymous nucleotide divergence (ω)	Li et al. (1985), Nei & Gojobori (1986)
		HKA	Degree of polymorphism within and between species at two or more loci	Hudson et al. (1987)
		MKT	Ratios of synonymous and non-synonymous nucleotide divergence and polymorphism	McDonald & Kreitman (1991)
		DoS	Direction of selection	Stoletzki & Eyre-Walker (2011)

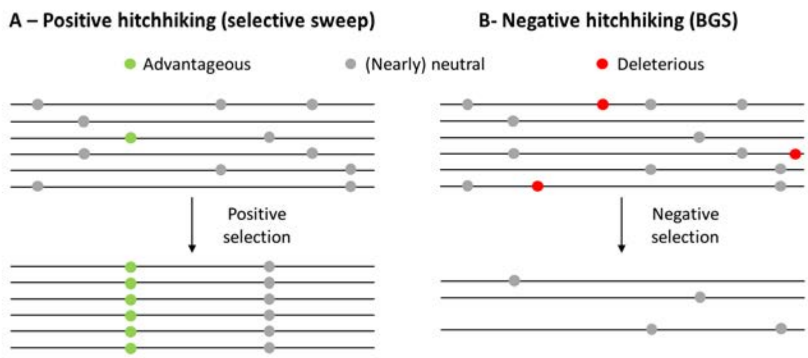


Figure 1.4: Effects of positive and negative hitchhiking models without recombination. (A) Positive hitchhiking model (selective sweep). In this case, an advantageous mutation (green) is positively selected and increases in frequency, along with linked (nearly) neutral mutations (grey). **(B)** Negative hitchhiking model (backgroud selection). Deleterious alleles (red) are removed from the population by puryfing selection, together with linked neutral variants. Both models predict a reduction of neutral variability. Note that without recombination, in the first scenario, all individuals from a population will show the same single haplotype containing the selected allele, whereas in the second one, several haplotypes will remain. (Adapted from Nachman 2001).

Tests based on the allele frequency spectrum and/or levels of variability

Several tests allow identifying regions in the genome which have been under positive selection pressures by analyzing the site frequency spectrum (SFS) of variants, looking for the surplus of rare derived alleles caused by selective sweeps. These tests compare the number of polymorphisms or derived singletons (variants observed only once) with the mean nucleotide diversity (Tajima's D and Fu and Li's F), the number of singletons with the total number of derived variants (Fu and Li's D) or the mutations at extreme low or high frequencies with the ones at intermediate frequencies (Fay and Wu's H).

Tests based on linkage disequilibrium

When an adaptive variant is positively selected, if its fixation process is faster than the reduction of LD between this selected mutation and linked variants by local recombination, the region will present long-range LD signals (Nielsen 2005, Franssen et al. 2015, Garud et al. 2015, Garud & Petrov 2016). Hence, certain unusually long haplotype will increase in frequency in the population. In this regard, several tests have been developed to detect the footprint of natural selection based on linkage disequilibrium metrics. These tests measure the decay of LD between loci with distance (EHH) or the frequency of alleles in regions of high LD (LRH, Hudson haplotype test, iHS), among others features.

Tests based on population differentiation

Selection acting on an allele in one population but not in another creates a marked difference in the frequency of that allele between the two populations. This effect of differentiation stands out against the differentiation between populations with respect to neutral (i.e., non-selected) alleles. Therefore, some neutrality tests analyze the alleles frequencies (XP-CLR), gene diversity (GST) or gene flow (F_{ST}) between subpopulations to identify putatively adaptive genome regions.

Tests based on comparisons of polymorphism and/or divergence between different classes of sites

The combined analysis of polymorphism and divergence rates is one of the most powerful approaches to understand the influence of different population genetics forces modeling the patterns of molecular evolutionary change, as they provide different but complementary information about the past evens experienced by any genomic region. In this regard, diverse tests that allow detecting positive selection signals in the genome based on comparisons of polymorphism and/or divergence between different classes of sites have been developed in the lasts years. Here, we focus on the two metrics which have been most used in population genomic studies: the K_A/K_S ratio and the McDonald and Kreitman test.

The K_A/K_S ratio

The K_A/K_S ratio, also known as ω (Yang & Bielawski 2000), compares the rate of non-synonymous divergence to the proportion of neutral fixations. Hence, it assumes one class of sites to be putatively adaptive while the other evolves under neutrality. This test was developed as an extension of the D_N/D_S ratio, which compares the total number of fixed sites (i.e., divergent) at the two same classes (Li et al. 1985, Nei & Gojobori 1986).

If advantageous mutations have been frequent in non-synonymous sites and have appeared at a higher frequency than do neutral ones, the rate of non-synonymous substitutions (K_A) will be greater than the rate of silent substitutions (K_S). Contrarily, if negative selection is constantly removing non-synonymous changes, K_A will be much lower than K_S. Thus, this ratio is used to indicate the functional constraint of any genomic region: $K_A/K_S = 1$ under neutrality, $K_A/K_S < 1$ under functional constraint, and $K_A/K_S > 1$ under positive selection.

This test is conservative because most non-synonymous mutations are expected to be deleterious. Thus, the proportion of adaptive substitutions has to be high for adaptive evolution to be detectable using this method.

The McDonald and Kreitman Test

The McDonald and Kreitman test (MKT, McDonald & Kreitman 1991) was developed as an extension of the Hudson–Kreitman–Aguadé test (Hudson et al. 1987) and is used to detect the signature of selection at the molecular level. Since its formulation, it has become one standard for testing adaptation in different classes of sites and organisms (Egea et al. 2008). The MKT compares the amount of variation within a species (polymorphism, P) to the divergence (D) between species at two types of sites, one of which is putatively neutral and used as the reference to detect selection at the other type of site.

In the standard MKT, these sites are synonymous (putatively neutral, 0) and non-synonymous sites (selected sites, i) in a coding region. Under strict neutrality, the ratio of the number of selected and neutral polymorphic sites (P_i/P_0) is equal to the ratio of the number of selected and neutral divergence sites (D_i/D_0). The null hypothesis of neutrality is rejected in a MKT when $D_i/D_0 \neq P_i/P_0$. The excess of divergence relative to polymorphism for class i, is interpreted as adaptive selection for a subset of sites i. The significance of the test can be assessed with a Fisher exact test (Fisher 1922) using the 2x2 contingency table.

An extension of the test allows to estimate the proportion of non-synonymous substitutions that have been fixed by positive selection (i.e., the rate of adaptive evolution, α; Charlesworth 1994, Smith & Eyre-Walker 2002). α values of 0 indicate that the region is under neutrality, $\alpha > 1$ is indicative of positive selection and finally, $\alpha < 0$ indicates negative or purifying selection on that region.

Tests derived from the MKT

In the standard McDonald and Kreitman test, the estimate of adaptive evolution (α) can be easily biased by the segregation of slightly deleterious non-synonymous substitutions (Eyre-Walker 2002) and different demographic histories. Specifically, slightly deleterious mutations contribute more to polymorphism than they do to divergence, and thus, lead to an underestimation of α in the cases where the population size has been relatively stable through time. On the contrary, if the population has experienced a process of expansion, some slightly deleterious mutations that were fixed in the past population with small N_e by genetic drift now contribute to divergence, and this produces an overestimation of α (Eyre-Walker 2002).

Several modifications to the standard MKT have been developed to address the most common situation, in which slightly deleterious mutations lead to an underestimation of the true level of adaptation. Here we review three of them: FWW method (Fay et al. 2001), DGRP correction (Mackay et al. 2012) and asymptotic MK method (Messer & Petrov 2013).

FWW correction method

Because slightly deleterious mutations tend to segregate at lower frequencies than do neutral mutations, they can be partially controlled by removing low frequency polymorphisms from the analysis. This is known as the FWW method (Fay et al. 2001). In this case, α is estimated considering only those polymorphic sites (for both neutral and selected classes) with a frequency above the arbitrarily established cutoff (x).

While this correction removes slightly deleterious mutations and therefore leads to α estimates significantly higher than the standard method, it also gets rid of all informative neutral polymorphism, which may cause a lack of statistical power to perform the test due to the low number of segregating sites remaining after trimming.

DGRP correction method

An alternative approach which also considers the presence of non-synonymous slightly deleterious mutations is the DGRP methodology (Mackay et al. 2012). Because adaptive mutations and weakly deleterious selection act in opposite directions on the MKT, α and the fraction of substitutions that are slightly deleterious will be both underestimated when both selection regimes occur. To take adaptive and slightly deleterious mutations mutually into account, P_i, the count of segregating sites in class i, is separated into the number of neutral variants and the number of weakly deleterious variants, $P_i = P_{i\ neutral} + P_{i\ weakly\ deleterious}$. Then, α is estimated using the standard MKT expression but substituting P_i by the expected number of neutral segregating sites, $P_{i\ neutral}$.

This method also permits quantifying the fraction of sites within the selected class under purifying selection. The excess of sites segregating with a frequency below the cut-off with respect to the neutral site class are considered to be weakly deleterious (b). Then, the neutral fraction (f) of putatively selected sites is estimated from the neutral class after correcting for weakly deleterious sites. Finally, the fraction of new mutants which are strongly deleterious and therefore not segregating (d) is estimated based on the previous fractions.

Asymptotic MK method

Messer & Petrov (2013) proposed a simple asymptotic extension of the MKT that yields accurate estimates of α, as it also takes slightly deleterious mutations presence into account (asymptotic MK method, Figure 1.5). Briefly, this approach first estimates α for each derived allele frequency (DAF) category using its specific P_i and P_0 counts. Because $\alpha_{(x)}$ depends only on the ratio $P_{i(x)}/P_{0(x)}$, any biases affecting the SFS at functional and synonymous sites in the same way, regardless whether due to demography or genetic draft, effectively cancel out. Then, the method fits an exponential function to this values and finally, the asymptotic α estimate (α_{asym}) is obtained by extrapolating the value of this function to 1, where it should converge close to the true α assuming that adaptive mutations do not significantly contribute to polymorphism and that purifying selection has been sufficiently stable over time.

While this test is able to properly overcome the presence of slightly deleterious variants and provides the most accurate estimates of α (Messer & Petrov 2013, Haller & Messer 2017), it has one major restriction which limits its application. To obtain unbiased α values the test requires a large amount of putatively selected segregating sites spanning the complete derived allele frequency spectrum, which is unrealistic for many *Drosophila* and human genes. Thus, this method is adequate for analyzing large genome regions or sets composed by multiple genes, but fails in the analysis of most single genes and small genomic regions with low to moderate variability.

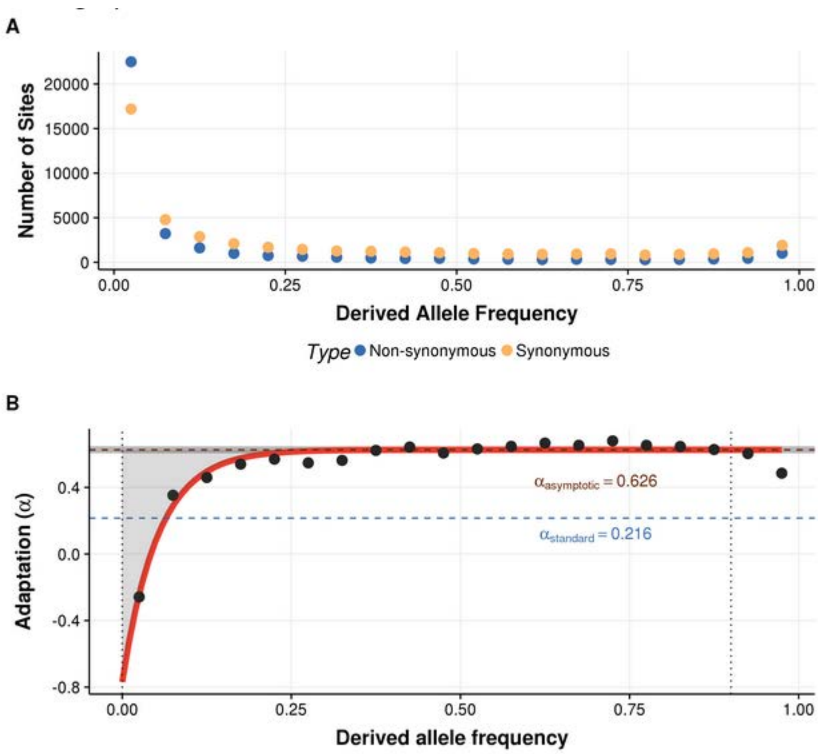


Figure 1.5: Derived allele frequency (DAF) and estimation of α by asymptotic MKT. (A) Number of non-synonymous and synonymous (neutral) polymorphic sites for each derived allele frequency. **(B)** Graphical representation of the estimation of α by the asymptotic MKT (Messer & Petrov 2013), showing (i) the α values for each DAF category (black dots), (ii) the model fitting (red line), and (iii) the α_{asym} and $\alpha_{original}$ values. Note that the estimate of α obtained with asymptotic MKT is substantially higher than the standard MKT one. Polymorphism and divergence values correspond to the complete chromosome arm 2R of a North American population (RAL, Raleigh, NC) of *D. melanogaster*, using *D. simulans* as outgroup species.

1.3 *Drosophila* as a model organism for population genetics

Fruit flies of the genus *Drosophila* have been an attractive and effective genetic model organism since their introduction as a research tool for biological studies in the early 20th century (Morgan 1915, Muller 1927). Thus, during the last 100 years, *Drosophila* has played a crucial role in the diverse fields of genetic analysis, such as ecology, speciation, developmental biology or population genetics (Powell 1997, Hales et al. 2015).

In the late 30s, Dobhansky and colleagues performed the firsts population genetic studies using *Drosophila*, focused on the analysis of inversions polymorphisms (Dobzhansky 1937, Dobzhansky & Sturtevant 1938). Few years later, *Drosophila* established as the most extensively used organism for studying genetic variation in natural populations (Lewontin & Hubby 1966, Ayala et al. 1974, Singh & Rhomberg 1987) and its importance within the population genetics field just increased more and more over time.

The complete *D. melanogaster* genome sequence was obtained in the late 90s, as part of one of the pioneers sequencing projects using whole-genome shotgun technologies in eukaryotic genomes (Rubin 1996, Adams et al. 2000). Specifically, it is around 180Mb, of which 120Mb is euchromatic genome and 75% of this fraction is intergenic or intronic (i.e., non-coding) sequence (Misra et al. 2002). It is divided in 5 chromosomes (X, Y, 2, 3, 4; Hoskins et al. 2002) and the current genome version includes 17,753 annotated genes, of which 13,931 are protein-coding genes (FlyBase release 6.22, August 2018).

Then, dozens of complete genomes started to become available for different species within the *Drosophila* genus (Clark et al. 2007). Nowadays, thanks to high-throughput sequencing technologies, hundreds of wild derived samples have been sequenced, and some of the most important population genetics projects in the last 10 years have been performed using *D. melanogaster*. Indeed, more than 1000 individual genome sequences of this species are available right now (Lack et al. 2016). In addition, new studies focused on other species of the genus *Drosophila* have been performed too. A great example is the recently published new panel of 170 inbred genotypes of a North American population of *D. simulans* by Signor et al. (2018), which will serve as a valuable complement to the DGRP and other *D. melanogaster* panels.

Overall, all these population genomics projects provide the fly lineage with a unique resource on which to test the molecular population genetics hypotheses and eventually understand the evolutionary dynamics of genetic variation in populations.

1.3.1 Population genomics projects in *Drosophila melanogaster*

Population genomics is understood as the study of the whole genomes of individuals instead of biased and fragmentary sequence samples. The first population genomics study in a *Drosophila* species was performed by Begun et al. (2007) in *D. simulans.* Authors analyzed genome-wide patterns of polymorphism and divergence (with *D. melanogaster* and *D. yakuba*) along the genome using 7 inbred lines of diverse origin. In addition, they also found evidence of adaptive protein evolution in the genome, showing for the first time that natural selection is pervasive in Drosophila.

Regarding *D. melanogaster*, the first study of natural variation was done by Sackton et al. (2009). In this work, nine strains from African (n=3) and North American (n=6) populations were analyzed and compared. After this preliminary study, two major population genomics-oriented projects using *D. melanogaster* and high-throughput sequencing technologies were developed:

- *The Drosophila Genetic Reference Panel (DGRP).* The DGRP, a community resource for the analysis of population genomics and quantitative traits, is a panel of > 200 inbred, mostly homozygous lines of *D. melanogaster* derived from a North American natural population (Raleigh, NC) (Mackay et al. 2012, Huang et al. 2014). This was the first population genomics study performed in *D. melanogaster*, and it allowed the description of genome-wide patterns of nucleotide diversity and divergence, and the mapping the diverse natural selection forces along the genome, as well as phenotypic analyses and the identification of several eQTLs.

- *The Drosophila Population Genomics Project (DPGP).* The DPGP (Langley et al. 2012) independently analyzed two natural populations of *D. melanogaster*: 37 DGRP lines and 6 lines from an African population of Malawi. The second freeze of the DPGP (DPGP2) sequenced the genomes of *D. melanogaster* from populations in Sub-Saharan Africa and France (Pool et al. 2012). The latest version of DPGP (DPGP3) involved the sequencing of around 200 individuals from a single population in Zambia, the region presumed to be the ancestral range of the species (Pool et al. 2012, Lack et al. 2015). These population data allowed the study of *D. melanogaster* demographic and migratory history (Pool et al. 2012), natural selection and the genetic basis of local adaptation (Langley et al. 2012) or chromosomal inversion polymorphism (Corbett-Detig & Hartl 2012) among others.

Besides, other population genetics projects in this species have been carried out in the last few years, providing even more available genome sequences: Campo et al. (2013), Kao et al. (2015), Bergman & Haddrill (2015), or the Global Diversity Lines published in Grenier et al. (2015).

Finally, the *Drosophila Genome Nexus* (DGN) is a recent compilation of each of these population genomic sequences assembled against a single common reference genome assembly. This project aims to increase the comparability of population genomic data sets, facilitating direct comparisons among them (Lack et al. 2015). In detail, DGN re-aligned genome sequences from: DPGP1 (Langley et al. 2012): 27 genomes from Malawi; DPGP2 (Pool et al. 2012): 139 genomes from 22 populations, mainly from Africa; DPGP3 (Lack et al. 2015): 197 genomes from Zambia; DGRP (Mackay et al. 2012, Huang et al. 2014): 205 genomes from Raleigh, USA; the global diversity lines (Grenier et al. 2015): 85 genomes from Australia, China, the Netherlands, the USA and Zimbabwe; Bergman & Haddrill (2015): 50 genomes from France, Ghana and the USA; Campo et al. (2013): 35 genomes from California; Kao et al. (2015): 23 genomes from 12 New World locations; and 306 new sequenced genomes from Ethiopia, South Africa, Egypt and France; resulting in a dataset of 1,067 complete sequence genomes which cover almost the complete geographical range of this species.

1.3.2 The *Drosophila Genome Nexus* sequence data

The *Drosophila Genome Nexus* project (DGN, Lack et al. 2015, 2016) provides the genome sequences of over 1,100 worldwide wild-derived *Drosophila melanogaster* individuals from 58 populations out of 23 countries spanning 5 continents (all but Antarctica).

DGN sequences were obtained following a two round assembly pipeline that combines two different aligners (Figure 1.6). Briefly, short reads were first mapped to the reference genome (version 5.57 from FlyBase) using BWA v0.5.9 (Li & Durbin 2010) and then the remaining unmapped reads were mapped using Stampy v1.0.20 (Lunter & Goodson 2011). All reads with mapping quality scores below the standard threshold were excluded. Then, optical duplicates were removed and assemblies were realigned around indels. After that, variants were called using GATK Unified Genotyper (DePristo et al. 2011), and the called SNPs and indels were inserted into the reference genome. This modified genome was used for the second round of mapping. Finally, base coordinates were shifted back to the original reference genome ones, in order to obtain aligned consensus sequences for all samples. Sites within 3' bp of a called indel and deletions were coded as "N". Insertions do not appear in the sequences. Therefore, DGN sequence data is highly appropriate for nucleotide variation analyses, but not for structural variation.

Moreover, there are some genomic regions which are especially problematic to analyze and must be taken with care and in some cases discarded when inferring major population genomics conclusions: regions that are classified as identical by descent (IBD), regions with genetic admixture, heterozygous regions that persist after many generations of full-sibling mating, or pseudo-heterozygous regions.

A genomic segment can be classified as IBD if two or more individuals have inherited it from a common ancestor without recombination; *i.e.,* the segment has the same ancestral origin in these individuals (Thompson 1975). Since most population genetic analyses assume that unrelated individuals have been sampled, it is important to

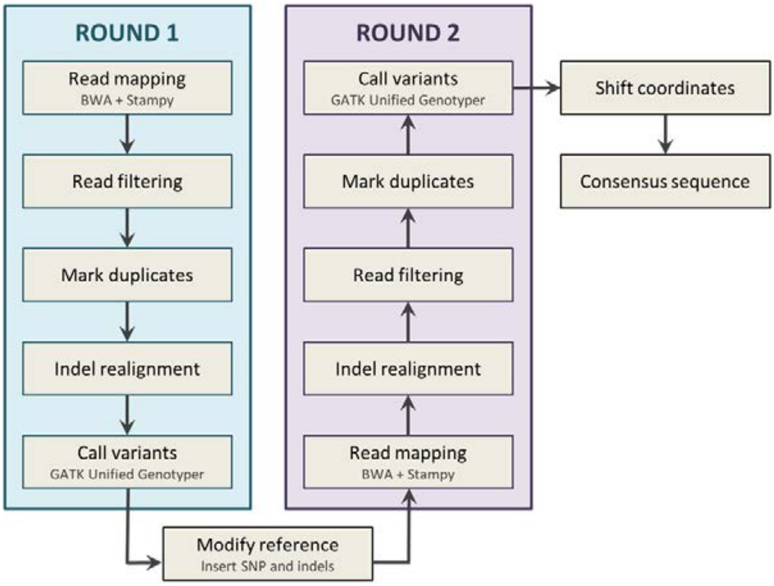


Figure 1.6: Drosophila Genome Nexus assembly pipeline. The pipeline consisting of two rounds of read mapping (using two different aligners), read filtering, indel realignment and calling of variants. Briefly, it takes sequencing reads as input and returns as output a consensus sequence for each sample, whose coordinates are shifted and aligned to the *D. melanogaster* reference version 5.57 (FlyBase). (Adapted from Lack et al. 2015)

take into account which regions are influenced by the effects of close relatedness among genomes. Besides, genetic admixture occurs when two or more isolated populations begin to interbreed. This ultimately causes the homogenization of such populations and the loss of the "unique" variation of each of them. The presence of admixture in both African and non-African populations has been documented in Pool et al. (2012).

Finally, heterozygosity can persist in fly stocks even after many generations of full-sibling mating, probably due to the presence of recessive lethal or infertile mutations, which are commonly found on wild-derived *Drosophila* chromosomes. In addition, while haploid embryo genomes are not expected to contain any true heterozygosity, repetitive and/or duplicated regions can cause mis-mapping that results in tracts of "pseudo-heterozygosity".

Thus, the genome sequences resulting from the DGN assembly pipeline were then filtered to remove the problematic regions from the alignments. Specifically, these regions were re-encoded as “N” in individual samples sequences (Lack et al. 2016), becoming non-informative for population genetics analyses.

1.4 Bioinformatics of genetics diversity

High-throughput sequencing technologies are allowing the deciphering of an explosive number of nucleotide sequences in a large number of genes and species, making available complete genome sequences of hundreds or even thousands of individuals for certain species (e.g., Lack et al. 2016 for *D. melanogaster*; Consortium et al. 2015 for human). Indeed, the rate at which genomes for new species and within species individuals are being sequenced continues to accelerate thanks to the advance in sequencing technologies which lower the cost of obtaining these data. This is clearly observable in two of the main portals of genetic data storage (GenBank and The Genomes Online Database, Figure 1.7).

The GenBank portal, the US National Institute of Health (NIH) genetic sequence database, stores all publicly available DNA sequences (Benson et al. 2013; `https://www.ncbi.nlm.nih.gov/genbank/`). From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months. In its last statistics release (February 2018), this database reported to store 253,630,708,098 bp and 207,040,555 sequences for GenBank projects and 2,608,532,210,351 bp and 564,286,852 sequences regarding whole-genome sequencing (WGS) projects (Figure 1.7A). Besides, the Genomes Online Database (GOLD, `https://gold.jgi.doe.gov/`) tracks genome sequencing projects. At the current date (February 2018) it includes 148,641 sequenced genomes of organisms, of which 1,627 are archaeal, 99,973 bacterial, 38,145 eukaryal and 8,896 virus (Figure 1.7B).

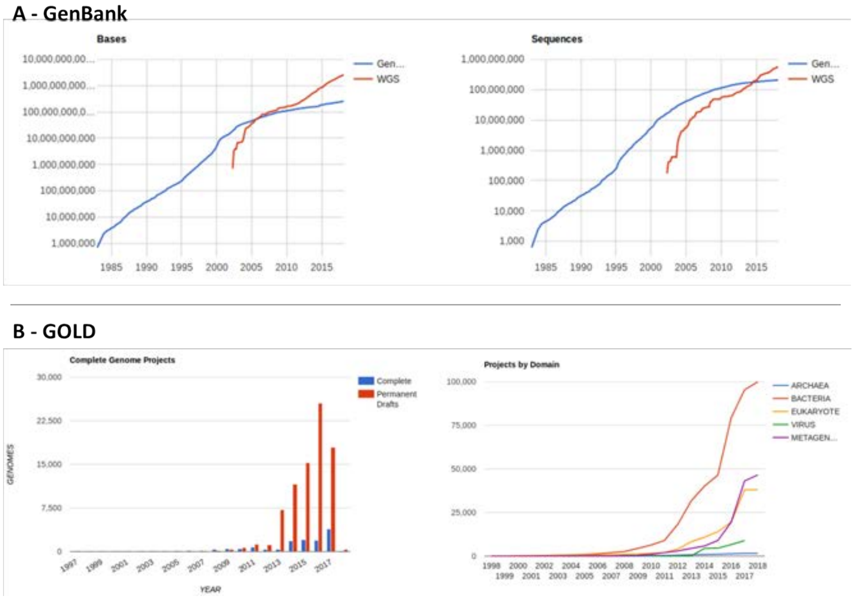


Figure 1.7: GenBank and GOLD storage metrics charts. (A): GenBank base pairs (left) and sequences (right) stored from 1985 to date (2018). The pattern follows a continuous increase along time even though it seems to accelerate at a lower rate in the last years. Note that values in the y axis increment by one order of magnitude and not linearly. In addition, the plot also shows the explosion of WGS projects from 2005, whose number of both bp and sequences is also increasing very quickly. (Retrieved from `https://www.ncbi.nlm.nih.gov/genbank/statistics/` in February 2018) **(B)** Genome Online Database (GOLD) complete genome projects (left) and number of projects classified by phylogenetic domains (right) from 1997 to 2018. The amount of complete sequenced genomes is also increasing exponentially over time. (Retrieved from `https://gold.jgi.doe.gov/statistics` in February 2018)

Bioinformatics, understood as he computational analysis of biological data in which biology, computer science and information technology merge into a single discipline, has become essential in order to facilitate the handling and analysis of such a huge amount of genetic information. Therefore, the development and implementation of bioinformatic tools addressing the scientific needs of all the steps from data acquisition, quality checking, and analysis, as well as storage and representation, has experienced an exponential increase in the last two decades; and this tendency does not seem to decelerate (Li et al. 2016).

1.4.1 Genome browsers: graphical biological databases

The initial major requirement arising from the explosion of genomic data is the development and establishment of molecular databases to store, organize and index these complex datasets.

A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query and retrieve components of the data stored within the system. Two major requirements are necessary for researchers to benefit from the data stored in a database: (i) easy access to the information, and (ii) a method for extracting only that information needed to answer a specific biological question (Casillas 2008).

Biological databases were classically grouped into broad subject categories, such as: (i) nucleic acid sequence and structure, and transcriptional regulation; (ii) protein sequence and structure; (iii) metabolic and signaling pathways, enzymes and networks; (iv) genomics of viruses, bacteria, protozoa and fungi; (v) genomics of human and model organisms plus comparative genomics; (vi) human genomic variation, diseases and drugs; (vii) plants and (viii) other topics, such as proteomics databases. However, as the scientific progress is moving towards an era of increasingly interdisciplinary research, the content of many databases may span multiple categories so that resources do not fill only within one specific category, but several at once, leaving obsolete the original classification schema. Overall, the number of new biological databases developed is constantly increasing over time, with more and more sites available on-line (Rigden & Fernández 2017).

Some initiatives are trying to integrate the major number of databases into general portals to facilitate the search of information in the current bioinformatics era. The most relevant examples are the European Bioinformatics Institute - EBI portal (`http://www.ebi.ac.uk`), and the National Center for Biotechnology Information - NCBI (`http://ncbi.nlm.nih.gov`).

Genome databases are a sub-category of molecular databases, specialized in storing and rendering genome-based data. Specifically, genome databases can be defined as data repositories, generally implemented via relational databases, that include the available genomic sequence data of one or more organisms, together with additional information, referred as annotations (Ràmia 2015). Thus, genome databases enable researches to formulate biological queries involving data originated from diverse sources (Schattner 2008) as they aggregate and integrate data from multiple databases in a uniform and standardized manner. This class of databases are also referred as secondary databases.

Within the genome databases category, one very specialized type are genome browsers. In general, these can be defined as visual secondary databases which provide a graphical interface for users to browse, search, retrieve and analyze genomic sequence and annotation data (Wang et al. 2013). Thus, usually they act as a central starting point for genomic research (Furey 2006).

Genome browsers are based on a reference genome, used as a coordinate system on which diverse tracks (i.e., non-overlapping layers of information covering any specific region of the genome) are graphically displayed. The essence of a genome browser is to pile up multiple tracks under the same genomic coordinates along the Y-axis, thus users could easily examine the consistency or difference of the annotation data and make their judgments of the functions or other features of the genomic region (Wang et al. 2013).

In general, genome browsers can be divided into: (i) stand-alone (or desktop) applications and (ii) web-based browsers. Stand-alone applications (such as IGV, Thorvaldsdóttir et al. 2013) provide an empty browser structure which has to be filled by the user. They include all the default functionalities of genome browsers to search, navigate and display diverse types of annotations, but no data at all. These applications are usually devoted to provide a fast and easy way to visualize specific projects' heterogeneous data in a unified manner, but do not allow sharing information through the web. On the other hand, web-based genome browsers are useful tools in promoting biological research due to their flexible (and often free) on-line accessibility and high performance.

There are two types of web-based genome browsers (Table 1.3). The first type is the multiple-species genome browsers, which integrate sequence and annotations for dozens of organisms and further promote cross-species comparative analysis. Most of them contain abundant annotations, covering gene model, transcript evidence, expression profiles, regulatory data, etc. The other type is the species-specific genome browsers which mainly focus on one model organism and may have more annotations for a particular species. The diverse species-specific genome browsers are powered by the Generic Model Organism Database project (GMOD, `http://www.gmod.org`). The GMOD Community is a bioinformatics initiative which aims to clarify and standardize genomic data analysis procedures, and provides a collection of open source software tools for managing, visualizing, storing, and disseminating genetic and genomic data. One of the most used tools from GMOD is the GBrowse framework (Stein et al. 2002). Indeed, most of the species-specific genome browsers are originally implemented using the GBrowse software, although since 2017 some of them are shifting to the newest browser framework developed by GMOD, named JBrowse (Skinner et al. 2009, Buels et al. 2016).

Thus, besides the implementation of curated web-based genome browsers, several genome browser frameworks have been developed in the last years (Table 1.4). These genome browser frameworks are empty architectures which have to be installed locally, configured and customized by users with their own genome annotation data. Then, they are shared through the web to make all data accessible and freely available. The use of well-designed genome browser frameworks allowed the development of specific genome browsers focused in certain areas within the broad genetics/genomics field.

Table 1.3: Web-based genome browsers. List of web-based genome browsers classified in multiple-species and species-specific platforms. All species-specific genome browsers listed here except MGI were initially based on the GBrowse framework developed by GMOD and now are shifting to JBrowse.

	Name	URL	Description	Reference
Multi-species	NCBI GDV	`https://www.ncbi.nlm.nih.gov/genome/gdv/`	Genome Data Viewer: major species with completed genome sequences including vertebrates, invertebrates, protozoa, plants and fungi, as well as dozens of uncompleted plant genomes	Coordinators (2016)
	UCSC	`http://genome.ucsc.edu/`	Major species with completed genome sequences including vertebrates, deuterostomes, insects and nematodes. No plant species	Karolchik et al. (2003)
	Ensmbl	`http://www.ensembl.org/index.html`	Major species with completed genome sequences providing lineage-specific web portals for vertebrates, metazoa, plants, fungi, protists and bacteria	Hubbard et al. (2002)
	VISTA	`http://pipeline.lbl.gov/cgi-bin/gateway2`	Whole genome alignment presentation, including vertebrates, insects, nematodes, deuterostomes, plants, fungi, alga, annelids, stramenopiles and metazoa	Poliakov et al. (2014)
Species-specific	FlyBase	`http://flybase.org/jbrowse/?data=data/json/dmel`	*Drosophila* (Fruit fly) genes and genomes information, including gene expression, mutations, microarray features, non-coding information, etc.	Gramates et al. (2016)
	WormBase	`http://www.wormbase.org/tools/genome/jbrowse-simple/?data=data/c_elegans_PRJNA13758`	*Caenorhaditus elegans* (Worm) complete genome information regarding genes, variation, genome structure, modENCODE data, and so on.	Harris et al. (2009)
	MGI	`http://jbrowse.informatics.jax.org/?data=data/mouse`	Mus musculus (Mouse) genes and genome sequence data, with a complete catalog of genotype-phenotype known associations	Smith et al. (2017)
	ZFIN	`http://zfin.org/action/gbrowse/`	*Danio rareo* (Zebrafish) genes, transcripts, phenotype, expression, knockdown reagent and mutation project information. Based on GBrowse.	Howe et al. (2012)

Table 1.4: Genome browser frameworks. List of the currently available genome browser frameworks on which most specialized genome browsers are based. (Adapted from Wang et al. 2013)

Name	URL	Description
Ensembl	`http://www.ensembl.org`	An extensible software architecture with powerful API support
UCSC	`http://genome.ucsc.edu`	A freely downloadable package for local browser installation
GBrowse	`http://gmod.org/wiki/GBrowse`	The most popular genome viewer for multiple genome annotation visualization, especially for model organism database projects
JBrowse	`http://jbrowse.org`	A JavaScript-based genome browser, providing Google-map like browsing experience
ABrowse	`http://www.abrowse.org`	A new-generation customizable genome browser framework
Anno-J	`http://www.annoj.org`	An interactive application designed for visualizing genome annotation data and deep sequencing data
SViewer	`http://www.ncbi.nlm.nih.gov/projects/sviewer`	NCBI's new sequence viewer

Some clear examples of population genetics-oriented genome browsers are PopDrowser (Ràmia et al. 2011), The 1000 genomes selection browser 1.0 (Pybus et al. 2013) and the *Drosophila buzzati* genome project browser (Guillén et al. 2014).

JBrowse software

The GBrowse genome browser framework developed by the GMOD community rapidly became the most powerful and used tool for custom-browsers development. However, this platform has become outdated in terms of performance and data display. Thus, GMOD developers presented a new generation genome browser framework named JBrowse. The first version was released in 2009 (Skinner et al. 2009), but it was not until 2016 that the stable software release was presented (Buels et al. 2016).

JBrowse can be defined as a fast, embeddable genome browser framework built completely with JavaScript and HTML5, with optional run-once data formatting tools written in Perl, that can be used to navigate genome annotations over the web. Compared to other existing genome browser frameworks (including GBrowse), JBrowse presents two clear advantages: (i) it helps preserve the user's sense of location by avoiding discontinuous transitions, instead offering smoothly animated panning, zooming, navigation, and track selection; and (ii) it distributes work between the server and the client, and most tracks are directly loaded on the client side (without using pre-rendered images), offering a better user experience and speeding up the browser performance (Skinner et al. 2009).

When a web browser loads a page containing JBrowse and creates a Browser object (the main controlling object for a JBrowse instance), the first thing the Browser does is to read the configuration information. Then, based on the configuration information, Jbrowse client decides (i) the reference sequence providing the coordinate system and sequence data for a given dataset and, (ii) the set of available annotation tracks which may be rendered alongside these reference sequences to display (Buels et al. 2016).

The core visual elements of a JBrowse track are sequence data, feature glyphs, and quantitative data. Specifically, JBrowse supports the following major types of tracks:

- *Sequence (FASTA) tracks.* The Sequence track displays forward and reverse strands of the reference sequences and six conceptual translation frames. JBrowse can load sequence data from FASTA files, indexed FASTA files, and pre-processed sequence data converted into JSON files.

- *Feature (GFF, BED, GenBank) tracks.* The two types of tracks currently available for displaying annotations from GFF or BED files are HTMLFeatures and CanvasFeatures. These tracks can display features with optional structured subfeatures and are ideal for displaying gene models (with component exons, introns, UTRs), transcript alignments, single-nucleotide polymorphisms (SNPs), transposons, repeats, etc. Each type of tracks has its own customization options.

- *Quantitative (Wiggle, BigWig) tracks.* Numerical data stored in Wiggle and BigWig files can be plotted using histograms (the Wiggle/XYPlot track) or heat maps (the Wiggle/Density track). JBrowse can load quantitative data directly from BigWig files stored on the server, with no need for preprocessing.

- *Alignment (BAM) tracks.* Three types of track are available for rendering the data in BAM files (reference-aligned reads): Alignments (a highly configurable track with customizable click behavior which renders reads as individual HTML elements), Alignments2 (a faster track, optimized for deep-coverage datasets, which renders reads directly onto an HTML Canvas element), and SNPCoverage (a track which dynamically calculates and visually highlights SNPs from BAM data, including nucleotide frequencies). BAM files used with JBrowse must be compressed and coordinate-sorted.

- *Variant (VCF) tracks.* VCF can be rendered using the HTMLVariants and CanvasVariants track classes, derivatives of HTMLFeatures and CanvasFeatures that are optimized for displaying the potential large amounts of detailed data that go along with each variant. The VCF files must be compressed and indexed.

In summary, JBrowse is a fully-featured genome browser that is capable of visualizing diverse types of genome-located data, kept in a variety of different data stores, and of interfacing to other client and server applications. In addition, it is highly cross-platform; releases are tested on Mozilla-based browsers (e.g., Firefox), WebKit browsers (e.g., Safari, Chrome) and Microsoft Internet Explorer, and on desktop and mobile platforms with touchscreen support (Buels et al. 2016).

PopDrowser: the population Drosophila browser

One major issue in population genetics studies is how to properly visualize the diverse parameters estimated all together in their genomic context. In this regard, a key technological contribution from the Freeze 1 of the DGRP project (Mackay et al. 2012) was the development and implementation of a population genomics web browser, named PopDrowser, to make all genetic information available to the scientific community.

Thus, PopDrowser (the Population Drosophila Browser, Ràmia et al. 2011, `http://popdrowser.uab.cat`, Figure 1.8) is a genome browser

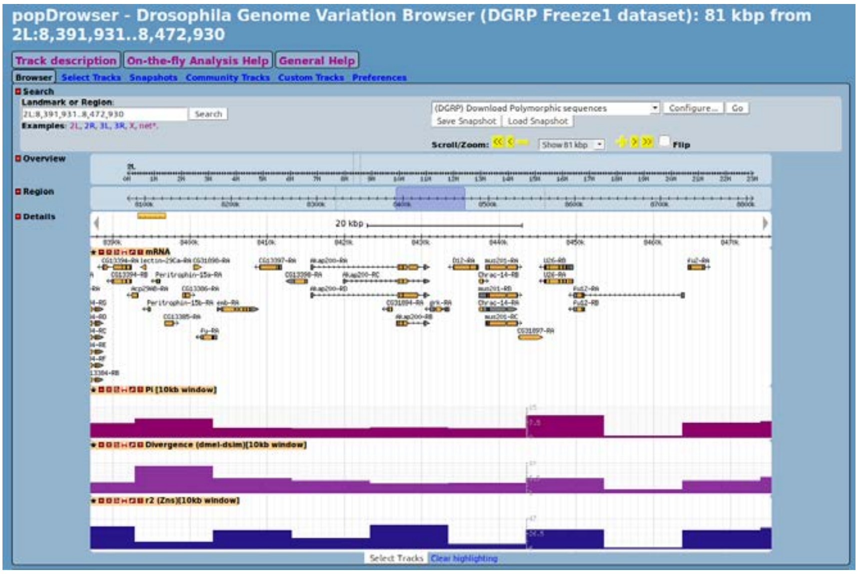


Figure 1.8: PopDrowser snapshot. PopDrowser snapshot showing the region *2L:8,391,931..8,472,930* with the following tracks activated: mRNA annotations, and nucleotide diversity (π), divergence with *D. simulans* (k) and r^2 metrics estimated in non-overlapping windows of 10 kb.

specially designed for the automatic analysis and representation of genetic variation metrics estimated from the DGRP data. Briefly, it allows reporting precomputed estimates of several DNA variation measures (polymorphism and divergence summary statistics, linkage disequilibrium parameters and several neutrality tests) along each chromosome arm through the combined implementation of the programs PDA2 (Casillas & Barbadilla 2006), MKT (Egea et al. 2008) and VariScan2 (Hutter et al. 2006) on a web-based user interface built on GBrowse software (Stein et al. 2002).

This browser was a pioneer tool in the field, becoming the first population genomics-oriented genome browser and storing the most extensive catalog of *D. melanogaster* population genetics metrics at that time (Mackay et al. 2012). In addition, it served for a variety of future studies in this model species (Koh et al. 2014, Reeves et al. 2014, Matute et al. 2014, Castellano et al. 2016, Salvador-Martínez et al. 2017).

1.4.2 The R project, an standard for statistical genetics analyses

Apart from the need for storing resources, another major requirement in the genomics era is to develop precise and specific statistical tools able to manage and analyze all this information. Over the last years, the open-source environment R (Ihaka & Gentleman 1996) has established as the most popular environment for statistical computing and data analysis across many fields of research, including genetics (Mair et al. 2015).

R can be defined as both, a language and an environment for statistical computing analysis (Team et al. 2013) that provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, while it also allows users to easily add additional functionality by defining new functions.

In addition, it also includes a suite of software facilities for data manipulation, calculation and graphical display, which include: (i) an effective data handling and storage facility, (ii) a suite of operators for calculations on arrays, in particular matrices, (iii) a large, coherent, integrated collection of intermediate tools for data analysis, (iv) graphical facilities for data analysis and display either on-screen or on hard-copy, and (v) a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities (Team et al. 2013).

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License (`https://www.gnu.org/licenses`) in source code form. It compiles and runs on the majority of operating systems, including UNIX, Windows and MacOS.

One of the cornerstones of the R system for statistical computing is the multitude of packages contributed by numerous package authors, which makes an extremely broad range of statistical techniques and other quantitative methods freely available (Mair et al. 2015). There are eight packages supplied with the R distribution and many more are available through communities like the Comprehensive R Archive Network (CRAN; `CRAN.R-project.org/`), Bioconductor

(Gentleman et al. 2004, `www.Bioconductor.org/`), R-Forge (Theußl & Zeileis 2009, `R-Forge.R-project.org/`), and GitHub (`https://github.com/`). The majority of R packages are released under open-source licenses, thereby placing no restrictions on users and guaranteeing that these packages can become public goods (Hippel & Krogh 2003).

In this regard, many R packages focused on genetics analyses have been developed over the last years. Some examples are: *diveRsity* (Keenan et al. 2013) and *PopGenome* (Pfeifer et al. 2014) for population genetics analyses; *phylotools* (Revell 2012) for phylogenetic comparative biology; or *gtx* (`https://cran.r-project.org/web/packages/gtx/`, 2013) for genetic association studies.

1.5 Objectives

This thesis is conceived as a comprehensive bioinformatics and population genetics project. It is centered in the development and application of bioinformatics tools for the analysis and visualization of nucleotide variation patterns and the detection of selective pressures in worldwide wild-derived *D. melanogaster* populations.

The main goal is accomplished in three sequential steps: (i) create a catalog of population genetics metrics and implement a tool for the graphical display of such information; (ii)develop a statistical package for the computation of the diverse selection regimes acting on genomes, and finally (iii) perform a global population genomics analysis in *D. melanogaster* using the previously developed tools.

Catalog of population genetics metrics and genome browser

First, create a complete inventory of population genetics parameters estimated both (i) in non-overlapping windows of varying size covering the complete euchromatic *D. melanogaster* genome and (ii) for each protein-coding gene, from the DGN sequence data (Lack et al. 2015, 2016). Then, implement a web-based genome browser open and freely available to allow the graphical visualization and retrieval of such information.

Statistical package for the computation of the diverse selection regimens acting on genomes

Second, develop an R package able to compute the McDonald and Kreitman test (McDonald & Kreitman 1991) and four tests derived from the MKT, using custom polymorphism and divergence genomic data. Provide ways to quantify both the rate of adaptive evolution (α) as well as the fraction of sites under purifying selection (d: strongly deleterious, b: weakly deleterious, f: neutral).

Population genomics analysis *in D. melanogaster*

Third, use the previously developed tools to perform an initial population genomics comparative analysis using populations of *D. melanogaster* of different geographical origin. Compare genome-wide polymorphism, divergence and historical recombination rates between populations; and quantify the fraction of sites under adaptive and purifying selection. Finally, identify and characterize the impact of recombination and coding density on the levels of nucleotide variation and adaptation along the *Drosophila* genome.

2. Materials and Methods

2. Materials and Methods

2.1 Data description

2.1.1 *D. melanogaster* genome sequences

D. melanogaster genome sequences from the *Drosophila Genome Nexus* project (DGN, Lack et al. 2015, 2016) were retrieved from `http://johnpool.net/genomes.html`. These sequences were aligned to the reference genome sequence version 5.57.

The initial set of DGN sequences (1,067 samples from 58 populations) were filtered by identity by descent, admixture, and heterozygosity (as described in Lack et al. 2016) and the problematic regions were encoded as "N". Only populations with at least 4 sampled genome sequences with $> 80\%$ of called alleles (non-"N") were considered for the analyses. The sequences for the selected populations were grouped in multiFASTA format files (*see Box 2.1 for details*) according to the population they belong to. The analyzed data comprises 966 genome sequences from 30 populations out of 18 countries spanning 5 continents (Table 2.1).

2.1.2 Geographic units of analysis

We performed a phylogenetic tree reconstruction which reveals the *D. melanogaster* populations structure using population differentiation metrics (F_{ST} values) which were estimated between pairs of 15 different populations by Lack et al. (2016). F_{ST} values were averaged across chromosome arms X, 2L, 2R, 3L and 3R, each of which was analyzed using inversion-free chromosomes only. The tree reconstruction was performed with the tree inference web-server T-rex (Boc et al. 2012, `http://www.trex.uqam.ca/`) using Neighbor-Joining reconstruction method (Saitou & Nei 1987) and the graphical display was done using MEGA6 (Molecular Evolutionary Genetics Analysis

Table 2.1: Analyzed *Drosophila melanogaster* populations. Information retrieved from Lack et al. 2016. Elevation values are in meters and n refers to the number of samples analyzed.

ID	Country	Locality	Continent	Date	Latitude	Longitude	Elevation	n
AUS	Australia	Sorell TAS	Oceania	2004	-42.77	147.56	18	18
CHB	China	Beijin	Asia	9/1992	39.91	116.41	52	15
CO	Cameroon	Oku	Africa	4/2004	6.25	10.43	2169	10
EA	Ethiopia	Gambella	Africa	12/2011	8.25	34.59	525	24
EB	Ethiopia	Bonga	Africa	12/2011	7.26	36.25	1725	5
ED	Ethiopia	Dodola	Africa	12/2008	6.98	39.18	2492	5
EF	Ethiopia	Fiche	Africa	12/2011	9.81	38.63	3070	69
EG	Egypt	Cairo	Africa	1/2011	30.1	31.32	25	32
ER	Ethiopia	D. Birhan	Africa	12/2011	9.68	39.53	2840	5
EZ	Ethiopia	Ziway	Africa	12/2008	7.93	38.72	1642	4
FR	France	Lyon	Europe	7/2010	45.77	4.86	175	96
GA	Gabon	Franceville	Africa	3/2002	-1.65	13.6	332	10
GU	Guinea	Dondé	Africa	6/2005	10.7	-12.25	801	5
KN	Kenya	Nyahururu	Africa	1/2009	0.04	36.36	2303	5
KR	Kenya	Marigat	Africa	1/2009	0.47	35.98	1062	4
MW	Malawi	-	Africa	2001	-	-	-	9
NG	Nigeria	Maiduguri	Africa	9/2004	11.85	13.16	295	6
NTH	Netherlands	Houten	Europe	1998	52.02	5.1	4	19
RAL	USA	Raleigh, NC	America	2003	35.76	-78.66	91	205
RG	Rwanda	Gikongoro	Africa	12/2008	-2.49	28.92	1927	27
SB	South Africa	Barkly East	Africa	12/2011	-30.97	27.59	1800	5
SD	South Africa	Dullstroom	Africa	12/2011	-25.42	30.1	2000	81
SF	South Africa	Fouriesburg	Africa	12/2011	-28.6	28.05	1800	5
SP	South Africa	Phalaborwa	Africa	7/2010	-23.94	31.14	350	37
UG	Uganda	Namulonge	Africa	4/2005	0.53	32.6	1134	4
UK	Uganda	Kisoro	Africa	1/2012	-1.28	29.69	1925	5
USI	USA	Ithaca, NY	America	2004	42.35	-76.57	344	19
USW	USA	Winters, CA	America	1998	38.53	-121.97	41	35
ZI	Zambia	Siavonga	Africa	7/2010	-16.54	28.72	530	197
ZS	Zimbabwe	Sengwa	Africa	9/1990	-18.16	28.22	865	5

Version 6.0) software (Tamura et al. 2013).

In order to provide a general view of nucleotide variation patterns along the whole genome and to facilitate the handling of such a heterogeneous dataset when performing population genetics analyses, we have analyzed not only single wild-derived populations but also *Drosophila* meta-populations, which are aggregations of geographically-related populations (Table 2.2).

Specifically, we considered 6 meta-populations: Asia, Oceania, America, Europe/North Africa, Equatorial Africa and Southern Africa. Asia and Oceania meta-populations correspond to the single sampled population from that continent (CHB, China, with 18 samples, and AUS, Australia, with 15 samples, respectively). The remaining four meta-populations include samples from 3 to 5 populations for which subsets of 10 individuals were sampled to build the corresponding

Table 2.2: Analyzed meta-populations of *D. melanogaster*.

ID	Name	n	Populations included (n)
CHB	Asia	15	CHB(15)
AUS	Oceania	18	AUS(18)
AM	America	30	RAL(10), USI(10), USW(10)
ENA	Europe/North Africa	30	EG(10), FR(10), NTH(10)
EQA	Equatorial Africa	50	CO(10), EA(10), EF(10), GA(10), RG(10)
SA	Southern Africa	30	SD(10), SP(10), ZI(10)

meta-population (e.g., the American meta-population -AM- contains 10 RAL, 10 USI, and 10 USW samples). Note that we considered European (FR, France, and NTH, The Netherlands) and Northern African (EG, Egypt) populations in the same meta-population, as suggested by Lack et al. (2016).

In summary, population genetics statistics were calculated for the 30 *D. melanogaster* wild-derived populations listed in Table 2.1 and for four additional meta-populations (Table 2.2).

2.1.3 Functional annotations and outgroup species

We used the *D. melanogaster* reference genome version 5.57 annotations file (in GFF3 format, *see Box 2.1 for details*), retrieved from FlyBase (`http://www.flybase.org`), to assess the functional class of each position in the genome (see below), and to estimate gene and coding density metrics. To do the latter, genes were mapped to non-overlapping sliding windows of varying size covering the euchromatic genome based on start coordinates overlap.

The genome sequences of *Drosophila yakuba* (version 1.3 from FlyBase; Clark et al. 2007) and *Drosophila simulans* (version 2.0 from FlyBase; Hu et al. 2013) are used as outgroup species. These sequences were retrieved aligned to the *D. melanogaster* reference genome version 5.57.

The divergence time between *D. melanogaster* and the outgroup species (Figure 2.1) is estimated to be: ~ 5 MYA with *D. yakuba* and ~ 2.5 MYA with *D. simulans*. Thus, the usage of each outgroup species has its own strengths and weaknesses. It is known that the estimates of the rate of adaptive evolution can be biased if the divergence time between two species is short. This bias appears because of three factors: (i) mis-attribution of polymorphisms to divergence; (ii) the contribution of ancestral polymorphism to divergence; and (iii) different rates of fixation of neutral and advantageous mutations (Keightley & Eyre-Walker 2012). Therefore, *D. yakuba* seems more appropriate than *D. simulans* for detecting genome-wide signals of selection due to its larger divergence time with *D. melanogaster*. However, the genome sequence of *D. simulans* has better coverage and quality, and hence, it allows analyzing regions of the genome on which there is not genomic information available for *D. yakuba* (Hu et al. 2013). In addition, Keightley & Eyre-Walker (2012) demonstrated that α estimates obtained using *D. melanogaster* and *D. simulans* would be inflated only by a small non-significant extent, as the branch length of each species to the common ancestor is greater than 10 N_e generations ($\sim$ 17 N_e).

Overall, to take advantage of the complementary information provided by each outgroup species, we have estimated population genetics parameters using both of them.

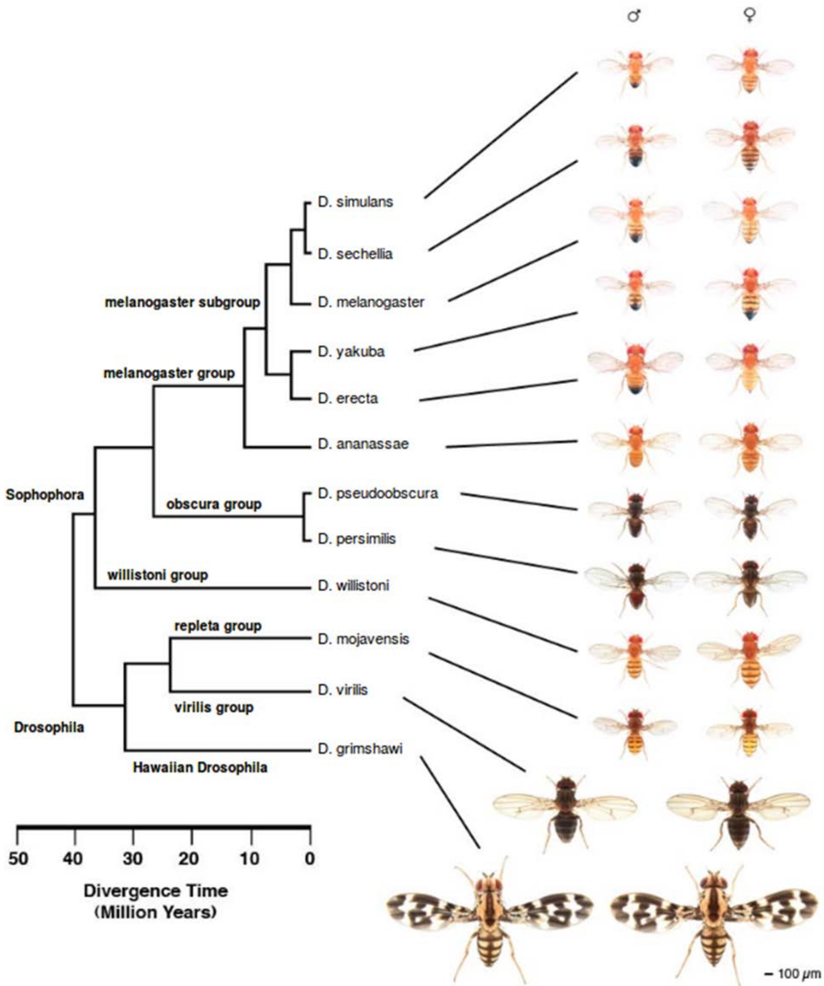


Figure 2.1: *Drosophila* phylogeny. Graphical representation of the phylogenetic relationship between *Drosophila* species along with images showing the male and female phenotypes. Note that all species used in this work are classified within the *melanogaster* group and subgroup. The divergence time between *D. melanogaster* and *D. yakuba* is ~ 3 times larger than the time between *D. melanogaster* and *D. simulans*. (Adapted from `http://flybase.org/static/sequenced_species` with images provided by Nicolas Gompel)

2.2 Estimation of population genetics parameters

We used DGN sequences, along with reference annotations and outgroup species data to generate a complete inventory of population genetics parameters, perform genome-wide analyses of nucleotide variation, and finally identify and assess the impact of the major genomic determinants of genetic variation and evolutionary rate.

Population genetics metrics were estimated using two different approaches: (*i*) non-overlapping sliding windows of varying size (1 kb, 10 kb, 50 kb, 100 kb) covering the *D. melanogaster* euchromatic genome (i.e., **windows-based** approach), and (*ii*) protein coding genes annotated in the reference genome (i.e., **genes-based** approach). Statistics were classified into 7 major categories: (i) frequency-based nucleotide variation, (ii) divergence-based metrics, (iii) linkage disequilibrium, (iv) historical recombination, (v) selection tests based on SFS and/or variability, (vi) selection tests based on polymorphism and divergence and, (vii) population differentiation.

2.2.1 Integration of annotations and sequence genomic data

Certain population genetics metrics, such as most selection tests based on polymorphism and divergence, depend on the comparison of the ratios of polymorphic and divergent sites at two functional classes of sites: neutral (0) and putatively selected (i).

Therefore, the integration of functional information into the genome sequences processing work-flow is required to estimate the number of analyzable (m_i, m_0), polymorphic (P_i, P_0) and divergent (D_i, D_0) putatively selected and neutral positions for each window and gene annotation. In this work, coding 4-fold degenerate sites were used as neutral reference and 0-fold degenerate sites as putatively selected. Coding 4-fold positions can present any nucleotide (A, C, G, T) without changing the amino acid of the translated protein and hence, represent the best proxy of truly neutral positions. On the other hand, 0-fold degenerated sites are restricted to one specific allele and any nucleotide change in such positions leads to an amino acid change in the resulting protein. Thus, they are assumed to be functional and putatively selected sites.

First, we selected the transcript with the highest number of coding positions (i.e., the longest CDS of the complete isoform) for each gene from the reference annotations file and we created temporary sequence files. Then, we produced 0-fold and 4-fold sequences which only encoded for that particular type of site, with the rest of positions coded as "N", using a custom Python script (Figure 2.2). The specific functional class of each coding position was assessed using the standard codon translation code table from NCBI (`https://www.ncbi.nlm.nih.gov/`). Briefly, we ensured that every CDS started with an "ATG" codon and finished with an STOP codon, taking into account the specific DNA chain and reading frame of each annotation. We followed a slightly different procedure for the windows-based and genes-based approaches:

- *Windows-based:* We first recoded the reference genome sequence. Then, we used the recoded reference as a template for recoding DGN samples and outgroup genome sequences (Figure 2.3). Finally, we grouped recoded samples files by population and used the resulting multiFASTA files for estimating population genetics metrics.

- *Genes-based:* We used DGN sequences, outgroup information and functional annotations data to create temporary multiFASTA sequence files for each gene and population, which where then recoded (Figure 2.5). Thus, with this approach we recoded the reference, DGN and outgroup sequences at the same time, removing from the analyses: (i) individual DGN samples which do not accomplish the recoding criteria and, (ii) non-homologous genes between *D. melanogaster* and the outgroup species.

Figure 2.2: Example of a recoded fragment of the longest isoform of a gene. First the CDS sequence is obtained, and then, using the degeneracy code, two sequences are created: 0-fold and 4-fold sites with the corresponding alleles and the rest of positions coded as "N". (Retrieved from `http://imk.uab.cat`)

2.2.2 Windows-based estimates

Table 2.3 includes the diverse population genetics metrics estimated using this approximation. Their estimation was implemented in an automated pipeline (Figure 2.3). The pipeline's input consists of (i) the aligned multiFASTA files containing DGN sequences grouped by population or meta-population, (ii) the reference *D. melanogaster* functional annotations (in GFF3 format) and (iii) genome sequence (in FASTA format) and, (iv) the outgroup reference genome sequences of *D. yakuba* and *D. simulans*. Then, population genetics metrics, classified in 7 categories, are calculated through the combination of custom ad hoc Perl and R scripts, Variscan2 (Hutter et al. 2006) and LDhelmet (Chan et al. 2012) software.

Frequency-based nucleotide variation

Genome sequences can include gaps (-) and missing or ambiguous nucleotides ("N"), which have to be discarded for the estimation of population genetics metrics. These problematic regions are not equally distributed along the genome and in all samples, and this might cause a bias in the estimation of statistics that where developed assuming constant sample size over the whole alignment. Thus, we did manually set the number of samples to use in order to ensure a constant sample size over all analyzed sites for each population.

Table 2.3: Windows-based population genetics metrics. List of major windows-based metrics, computed for each population and meta-population separately.

Category	Estimate	Description	Reference
Frequency-based nucleotide variation	S	Number of segregating sites per DNA sequence	Nei (1987)
	η	Total (minimum) number of mutations per site	Tajima (1996)
	η_E	Number of sites containing singletons	Tajima (1996)
	θ	Proportion of nucleotide sites that are expected to be polymorphic in any suitable sample	Watterson (1975); Tajima (1993, 1996)
	π	Nucleotide diversity: average number of nucleotide differences per site between each pair of sequences	Jukes & Cantor (1969); Nei & Gojobori (1986); Nei (1987)
	P_{0f}	Number of 0-fold (non-synonymous) segregating sites*	Nei (1987)
	P_{4f}	Number of 4-fold (synonymous) segregating sites*	Nei (1987)
	π_{0f}	Nucleotide diversity per bp in 0-fold sites*	Jukes & Cantor (1969); Nei & Gojobori (1986); Nei (1987)
	π_{4f}	Nucleotide diversity per bp in 4-fold sites*	Jukes & Cantor (1969); Nei & Gojobori (1986); Nei (1987)
Divergence-based metrics	K	Nucleotide divergence per bp,corrected by Jukes-Cantor	Jukes & Cantor (1969)
	D_{0f}	Number of 0-fold (non-synonymous) divergent sites*	Jukes & Cantor (1969)
	D_{4f}	Number of 4-fold (synonymous) divergent sites*	Jukes & Cantor (1969)
	K_{0f}	Nucleotide divergence per bp in 0-fold sites*	Jukes & Cantor (1969)
	K_{4f}	Nucleotide divergence per bp in 4-fold sites*	Jukes & Cantor (1969)
Linkage Disequilibrium	D	Coefficient of LD whose range depends of the allele frequencies	Lewontin & Kojima (1960)
	D'	Normalized D, independent of allele frequencies	Lewontin (1964)
	$\|D\|$	Absolute D value ($\|D\|$) averaged over all comparisons in the window	Lewontin & Kojima (1960)
	$\|D'\|$	Absolute D' value ($\|D'\|$) averaged over all comparisons in the window	Lewontin (1964)
	r^2	Statistical correlation between pairs of sites	Hill & Robertson (1968)
	h	Number of haplotypes	Nei (1987)
	Hd	Haplotype diversity	Nei (1987)
Historical recombination	ρ	Historical population-scaled recombination rate ($\rho = 4\ N_e\ r$)*	Chan et al. (2012)
Selection tests based on SFS and/or variability	Fu and Li's D	Number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants	Fu & Li (1993)
	Fu and Li's F	Number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences	Fu & Li (1993)
	Tajima's D	Difference between the number of segregating sites and the average number of nucleotide differences	Tajima (1989)
	Fu's F_S	Test based on the allele frequency spectrum	Fu (1997)
Selection tests based on polymorphism and divergence	K_A/K_S	Ratio of non-synonymous to synonymous nucleotide divergence (ω)*	Nei & Gojobori (1986)
	NI	Neutrality index, which summarizes the four values in an MK test table as a ratio of ratios*	McDonald & Kreitman (1991)
	DoS	Direction of Selection: difference between the proportion of non-synonymous divergence and non-synonymous polymorphism*	Stoletzki & Eyre-Walker (2011)
	α	Proportion of adaptive substitutions from McDonald-Kreitman test (MKT)*	McDonald & Kreitman (1991); Charlesworth (1994); Smith & Eyre-Walker (2002)
Population differentiation	F_{ST}	Average levels of gene flow based on allele frequencies, under the infinite-sites model	Hudson et al. (1992)

* Only computed for populations with $n \geq 10$ and for 100 kb, 50 kb and 10 kb sliding windows

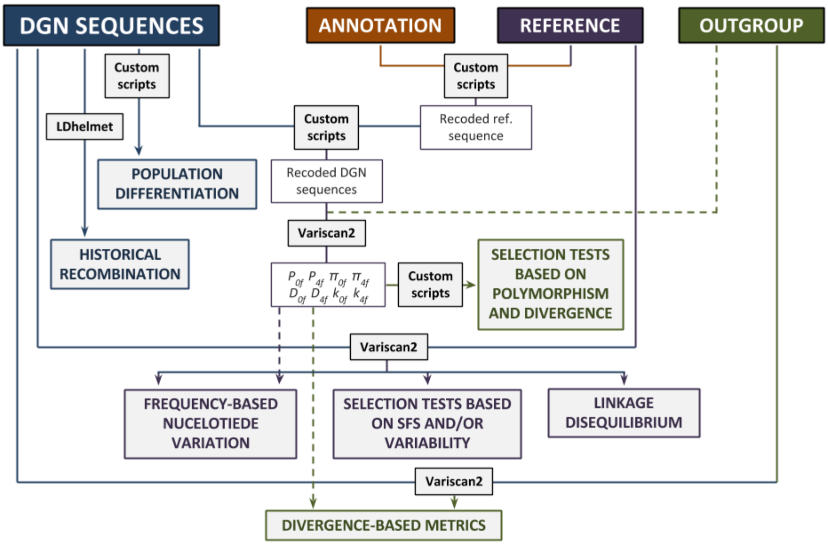


Figure 2.3: Windows-based approximation pipeline. This workflow allows the computation of several population genetics estimates for the DGN sequence data in an automated way, using a combined implementation of custom scripts, Variscan2 and LDhelmet software. Briefly, it takes as input the DGN sequences, reference sequence and functional annotations and outgroup sequence data, and returns population genetics metrics classified in 7 categories.

The number of samples depends on the sequences quality (understood as the proportion of called alleles relative to "N" calls), and ranges from 50% of individuals for populations with a high number of samples (e.g., RAL, ZI) to 100% for populations with a very low sample size (e.g., EB, ED, KN or UG). If a specific position contains less valid nucleotides than the established threshold, this position was excluded from analysis. On the contrary, if a site has a larger sample size than the threshold, extra nucleotides were randomly discarded. Figure 2.4 shows an illustrative example on how the handling of gaps and missing data is performed, setting the number of samples to use to ensure a constant sample size along the genome.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Seq 1	C	N	A	A	A	G	A	G	C	A	T	A	A	A	A	A	A	A	A	A
Seq 2	A	A	T	A	A	A	N	A	C	G	A	–	–	A	A	A	A	A	N	A
Seq 3	A	A	T	–	A	A	–	A	T	A	A	–	–	T	A	C	A	A	N	A
Seq 4	A	A	A	–	A	A	A	A	A	A	A	A	–	T	A	A	A	A	N	A
Seq 5	C	A	A	–	A	G	A	C	A	A	A	G	–	A	A	A	C	A	A	A
Seq 6	A	A	T	A	A	A	A	A	A	G	A	A	A	T	A	A	C	A	A	A
				*									*						*	

Figure 2.4: Handling gaps and missing data. In this example the minimum number of samples is set at four. Sites 4, 13 and 19 are discarded as they contain less than 4 valid nucleotides. Site 1 has 6 valid nucleotides, so 2 of those nucleotides will be randomly discarded. Site 2 contains one "N" which is automatically discarded. Out of the remaining 5 valid nucleotides 1 is randomly discarded to ensure the fixed sample size of 4. This procedure is repeated for all sites in the alignment. Since this leads to constant sample size over all analyzed sites, the standard statistics can be calculated. This feature is especially useful if e.g., only one sequence contains a long stretch of "N" or gaps, while all other sequences carry valid nucleotides. (Adapted from VariScan User's Guide v. 2.0, Hutter et al. 2006).

Divergence-based metrics

Statistics were estimated using both *D. simulans* and *D. yakuba* as outgroup species. Metrics regarding each outgroup were computed separately. We followed the procedure described in the previous section for setting the number of analyzable samples and sites for each population.

Linkage Disequilibrium

Linkage Disequilibrium (LD) metrics can be estimated with or without considering outgroup sequence information. Without an outgroup the derived and ancestral state of a polymorphic site are inferred by the frequency of the segregating alleles. The major allele is set to the ancestral and the minor allele to the derived state. With this definition the coupling and repulsion phases for each pair-wise comparison are calculated. However, when an outgroup is defined it is possible to infer the ancestral state of a polymorphic site by using the outgroup for comparison. If it is not possible to infer the ancestral state by comparison to the outgroup, the site will be discarded for analysis. Since this might happen quite frequently, the number of sites used for calculating LD statistics is usually higher when not using an outgroup, even though the inclusion of an outgroup sequence adds more information to the analysis. In this work, LD metrics were estimated taking into account outgroup species information.

Only polymorphic sites that contain exactly two variants were used. All sites containing gaps (-), missing or ambiguous nucleotides were excluded from the analyses. Finally, sites containing 3 or more variants were ignored for LD analyses but used when calculating haplotype (h: number of haplotypes; Hd: haplotype diversity) and Fu's F_S statistics.

Thes LD statistics were not computed for America, Europe/North Africa, Equatorial Africa and Southern Africa meta-populations because they are aggregations of wild-derived populations. As only bi-allelic polymorphisms are used, and gaps, missing or ambiguous nucleotides are discarded, samples from these meta-populations do not share enough analyzable polymorphic sites in order to obtain confident LD estimates.

Historical recombination rate

Population-scaled historical recombination rates were computed using LDhelmet software (Chan et al. 2012). Briefly, this method allows estimating fine-scale recombination rates genome-wide from patterns of genetic variation employing a reversible-jump Markov Chain Monte Carlo (rjMCMC) mechanism. It also uses custom quadra-allelic mutation models and mutation rates, along with information from the available genomes of outgroup species to infer a distribution of the ancestral allele at each polymorphic site.

Recombination in *D. melanogaster* occurs only in females and therefore, there are only 2 sets of recombinant chromosomes in each generation. In addition, males are homozygous for the X chromosome (X0) and thus, the effective population size of the X chromosome must be scaled by 4/3 to be equivalent to the N_e of the autosomes. Overall, the population-scaled recombination rate between a pair of sites in the X chromosome is defined as (Equation 2.1):

$$\rho_X = 8/3\ N_e^X\ r_f^X \tag{2.1}$$

where N_e^X is the effective population size for X and r_f^X is the probability of recombination between the sites per generation per X chromosome in females. On the other hand, the population-scaled recombination rate between a pair of sites in an autosome is defined as (Equation 2.2):

$$\rho_A = 2\ N_e^A\ r_f^A \tag{2.2}$$

where N_e^A is the effective population size for the autosome and r_f^A is the recombination rate between the sites per generation per autosome in females.

The computational method used for the estimation of ρ is based on the levels of linkage disequilibrium observed in the sample sequences, which are known to vary depending on the number of analyzed samples. So, in order to obtain comparable error measures between populations that differ in the number of sampled genomes, while ensuring an optimal computational performance, ρ was estimated using samples of 10 individuals taken from each population. The 10 genome sequences for each population were selected among those with the lowest percentage of ambiguous or missing nucleotides, in order to obtain the most accurate estimation in each case and to avoid biases caused by differences in sequences quality.

It is known that inversions in heterozygosis can repress recombination between the affected genomic region (Kirkpatrick 2010). Polymorphic inversions identified in *D. melanogaster* span from 3 to 14 Mb for autosomic inversions and from 1.7 to 6 Mb for X chromosome inversions (Corbett-Detig & Hartl 2012, Huang et al. 2014). Hence, inversion polymorphisms have to be taken into account to avoid introducing strong biases in the overall population recombination rates estimates depending on the samples used for calculation. We ensured that subset populations reflect the distribution of inversions genotype frequencies from the original populations (Fisher Exact test p-value > 0.05 in all cases, using counts of standard and inverted genotype samples for original and subset populations). The inversion genotypes for individual samples were estimated by Corbett-Detig and retrieved from `www.johnpool.net/Updated_Inversions.xls`.

We used LDhelmet software with the following parameters: 1,000,000 iterations of computation after 100,000 iterations of burn-in, specific mutation rates (θ in units of 1/bp) for each population (distinguishing autosomes and X chromosome values), the mutation matrix with nucleotide transition probabilities used in Chan et al. (2012), and default values for the rest of parameters (block penalty: 50; grid of p-values: 0, 0.1, 10.0, 100; and number of Padé coefficients: 11).

Metrics were estimated in blocks of 50 SNPs, and the output of LDhelmet consists of a list of SNPs with their associated ρ estimates. Thus, to obtain average recombination estimates for each window, SNP-based metrics were mapped to non-overlapping windows of varying size (10, 50 and 100 kb) using custom R scripts.

Selection tests based on SFS and/or variability

The calculation of $Tajima's\ D$, $Fu\ and\ Li's\ F$ and $Fu\ and\ Li's\ D$ statistics was performed following the procedure described in the Frequency-based nucleotide variation section (see above) for setting the number of analyzable samples and sites for each population. Instead, $Fu's\ F_S$ values were estimated considering same sites as for LD metrics.

Selection tests based on polymorphism and divergence

These statistics were computed using metrics from Frequency-based nucleotide variation and Divergence-based metrics categories in neutral (4-fold) and putatively selected (0-fold) sites. Specifically, NI and Fisher's p-value were estimated considering the number of polymorphic (P) and divergent (D) sites, whereas K_A/K_S and DoS metrics were computed using relative frequencies (π, K). The rate of adaptive evolution (α) from the McDonald and Kreitman test (MKT) was estimated using both, the counts of sites and their relative frequencies.

Population differentiation

Population differentiation metrics were estimated using custom scripts developed by Dr. John E. Pool (from Madison-Wisconsin University). Specifically, F_{ST} estimators of Hudson et al. (1992) were computed.

Only those sites for which 50% of genomes have a called allele from the largest African population sample (ZI population, $n = 197$) were analyzed. For each site, that involves comparing all pairs of individual genomes for each pair of populations, only valid (non-"N") called sites were considered. By excluding missing data on a per-individual-allele basis, this implementation weights sites based on how much missing data they have. Only windows with more than 100 usable sites are reported. In total, 35 comparisons for each window size were performed. This includes 10 comparisons among the populations with more than 50 individuals (EF, FR, RAL, SD and ZI), and the 25 comparisons that involve one of those populations and another population with at least 20 individuals (EA, EG, RG, SP and USW).

2.2.3 Genes-based estimates

Gene-based population genetics statistics were computed with the aim of testing adaptation at the gene level. Therefore, we only analyzed putatively selected (0-fold, i) and neutral (4-fold, 0) positions following this approach. We used data regarding 13,753 protein coding genes annotated in the *D. melanogaster* reference genome. Metrics were computed for the 20 DGN populations and meta-populations with more than 10 sampled genomes (Table 2.1, Table 2.2). We used *D. simulans* and *D. yakuba* genome sequences and annotations as outgroup species to estimate derived allele frequencies and divergence metrics.

Figure 2.5 illustrates the main processing work-flow. It takes as input DGN and outgroup sequences, along with gene annotations information, and returns population genetics parameters estimated at the gene level classified in 3 different categories. The pipeline uses custom Python and Perl scripts for the re-encoding of sequences and the computation of frequency-based nucleotide variation and divergence-based metrics. Specifically, the custom Perl script estimates the number of analyzed, divergent and polymorphic (along with the corresponding derived allele frequencies) sites. "N" calls were excluded from the analyses. In detail, we considered a site as analyzed if it had at least one called allele (non "N") in any sample and the outgroup sequence (Figure 2.6).

Frequency-based nucleotide variation

Metrics in this category include the number of putatively selected (P_{0f}) and neutral (P_{4f}) polymorphic sites with their corresponding derived allele frequencies (DAF) spectrum.

We estimated DAF grouping polymorphic sites in 10 and 20 frequency categories (DAF10 and DAF20, respectively). We considered a site as polymorphic if it had at least two different called alleles in the in-group sequences. To estimate the derived allele frequencies, we considered only polymorphic sites in which one of the called alleles was the same in in-group and outgroup sequences, counting the alternative alleles as derived (Figure 2.6).

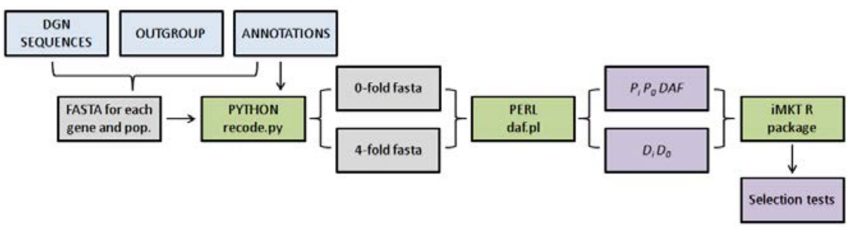


Figure 2.5: Genes-based approximation pipeline. This work-flow allows the computation of gene-based population genetics metrics. It takes as input DGN and outgroup sequences, along with reference annotations, and returns statistics from frequency-based nucleotide variation, divergence-based metrics and selection tests estimates. The main process is performed through a combination of Python and Perl ad hoc scripts.

Divergence-based metrics

We estimated the number of divergent positions for both classes of sites (D_{0f}, D_{4f}) considering a site as divergent if all alleles from the in-group sequences had the same called allele and it was different from the outgroup one (Figure 2.6).

Recombination

We used two different recombination estimates (which were computed at the window level) to assess gene-associated recombination values. Crossing-over meiotic recombination rates retrieved from Comeron et al. (2012) and historical population-scaled recombination metrics estimated using LDhelmet (Chan et al. 2012). Gene annotations were mapped to non-overlapping sliding windows based on start coordinates.

	1	2	3	4	5	6	7	8
Ingroup 1	A	A	A	A	A	A	A	A
Ingroup 2	A	A	A	T	T	A	A	A
Ingroup 3	A	A	T	T	C	A	A	T
Outgroup	A	C	A	A	A	—	N	N
Results	*	* *D*	* P (1/3)	* *P* (2/3)	* *P* (1/3 + 1/3)	—	—	—

Figure 2.6: Example of how to estimate the number of analyzable, polymorphic and divergent sites. The figure shows three sample in-group sequences together with outgroup sequence information and the assignment of each genomic position. In-group sequences were previously filtered as described in Figure 2.4. In total, 5 sites are analyzable (m), of which 1 is divergent (D) and 3 are polymorphic (P) with 4 different allele variants. The derived allele frequency of each polymorphic site is also shown. (**: analyzable position; – : non-analyzable position; D: divergent site; P: polymorphic site (with the frequency of derived alleles)*).

Selection tests based on polymorphism and divergence

We applied the Integrative McDonald and Kreitman Test (iMKT) R package (*see below*) to estimate diverse selection tests based on polymorphism and divergence genomic data for each analyzed gene and population. Specifically, we computed the following statistics: non-synonymous to synonymous divergence ratio (K_A/K_S), neutrality index (NI), direction of selection (DoS) and, ratio of adaptive evolution (α) using the original MKT and 4 derived methods: FWW correction (Fay et al. 2001), DGRP correction (Mackay et al. 2012), asymptotic MKT (Messer & Petrov 2013), and iMKT.

2.2.4 Statistical analyses

The statistical analyses shown in the section *3.3. Population genomics of D. melanogaster* were performed using the R software environment (`https://www.r-project.org`).

First, the significance of MKT-derived methods that are based on 2x2 contingency tables was assessed using a Fisher exact test (Fisher 1922) performed with the *fisher.test* function from the package *stats*.

Along this work, we also computed two different correlation coefficients with their corresponding p-values using the *cor.test* function from the package *stats*: (i) the Pearson's correlation coefficient (Pearson 1895), and (ii) the Spearman's rank correlation coefficient (Spearman 1904), using in both cases a significance level of 0.05.

Finally, we performed pair-wise Wilcoxon's rank sum tests (Wilcoxon 1945) in order to asses whether the differences in the nucleotide variation levels found between meta-populations and between the chromosome arms of each meta-population were statistically significant or not. Briefly, the Wilcoxon's signed-rank test is a non-parametric statistical hypothesis test used to compare two related samples to assess if their population mean ranks differ (i.e., it is a paired difference test) used when the data can not be assumed to be normally distributed (which is the case for population genetics data values such as π, k and ρ). The associated p-values were corrected using Holm's method (Holm 1979) in order to counteract the problem of multiple comparisons (the more hypotheses tested, the higher the probability of a Type I error, false positive). This approach controls the family-wise error rate (the probability that one or more Type I errors will occur) by adjusting the rejection criteria of each of the individual hypotheses or comparisons.

2.3 PopFly: the *Drosophila* population genomics browser

PopFly is a population genomics-oriented genome browser that allows the graphical display of a complete inventory of population genetics parameters estimated from the *Drosophila Genome Nexus* data.

2.3.1 Browser software and interface

The PopFly genome browser is build-up using JBrowse development version 12.1.1 software (March 2016, `http://jbrowse.org/jbrowse-1-12-1/`, Skinner et al. 2009, Buels et al. 2016).

The browser interface is coded mainly in JavaScript (JS) and HTML5, together with CSS for styling customization. The main configuration files of the browser are written in JS and JavaScript Object Notation (JSON). The first one controls the main features of the browser, and holds the structure and links among the diverse modules, while the second one contains detailed descriptions of all tracks stored (data location, meta-information, key, name, etc.). Both files are highly customizable by the administrator of the site.

Genetic information is stored using 4 major file formats (Box 1). Reference genome and DGN sequences are kept in both FASTA (multiFASTA) and VCF formats; gene, transposable elements and polymorphic inversions annotations are saved in GFF format and all functional and population genetics metrics estimated in non-overlapping sliding windows are stored in Wig or bigWig formats.

The current browser implementation is running under Apache on a CentOS 7 Linux x64 server, 16 Intel Xeon 2.4GHz processors, 32GB RAM.

2.3.2 Development and implementation of new utilities

Besides the default built-in functions of JBrowse, we have developed and implemented into the PopFly framework three new resources to facilitate population genetics data analyses and retrieval: (i) a tool designed to perform on-the-fly statistical analyses of the data, (ii) an application for displaying dynamic gene reports with information on adaptation metrics and, (iii) a plugin to download sequence information.

On-the-fly statistical analyses

The On-the-fly statistical analyses tool allows generating custom interactive plots to explore correlations among different genomic and geographical features estimated for each population.

This application was developed using the Shinyapps framework for R (`http://shinyapps.io`). Genetic population metrics correspond to 100 kb non-overlapping windows estimates, whereas geographical features were retrieved from Lack et al. (2016). The two correlation statistics provided (Pearson correlation coefficient, Pearson 1895, and Spearman's rank correlation coefficient, ρ, Spearman 1904) with their corresponding p-values are computed using a the *cor.test* function from R package stats, with default parameters. Finally, *ggplot2* and *shinythemes* R libraries are required for the graphical representation.

Gene adaptation metrics dynamic report

This function allows displaying a summary of adaptation metrics regarding any gene and population(s) of interest. In detail, the tool is accessible for 13,755 protein-coding genes annotated in the reference genome and 20 populations/meta-populations.

The map showing the available populations was built using Infogram software (`https://infogram.com/`). The dynamic report is generated using *rmarkdown* package (`https://cran.r-project.org/web/packages/rmarkdown/`) and implemented into PopFly through a combination of FastR framework (`https://github.com/oracle/fastr`) and custom JavaScript functions.

Download sequences

The download sequences plugin allows the retrieval of the genome sequences (in either FASTA or VCF file formats) corresponding to any region and population(s) of interest. This utility uses a combination of custom Perl and Bash scripts and vcftools software (Danecek et al. 2011) to perform the trimming and subset of the sequences of interest, along with JavaScript custom functions to implement the resource into the PopFly framework and manage the user's custom parameters request.

In addition to the plugin itself, we have also developed and implemented a *Cron Job* (i.e., a scheduled task that is executed by the system at a specified time/date) whose main functions are to ensure that there is enough free disk space in the server to use the utility (and if it is not the case, to provide a warning and a brief explanation about the problem) and to remove old files which are not necessary anymore, as they have been downloaded by the requesting user. This program is coded mainly using Bash programming.

Box 2.1: Genome annotations file formats

FASTA and multiFASTA: text-based format for representing raw sequences (Figure 2.7). The first line in a FASTA file starts with a greater than (>) symbol and holds a unique description of the sequence. Then, the following lines contain the sequence itself in standard one-letter code (Pearson & Lipman 1988). A multiple sequence FASTA (multiFASTA) format file is obtained by concatenating several single sequence FASTA files, with their corresponding identification and sequence lines.

```
>RAL-129_Chr2L:5020221_5045940
AAATAAAATCCAATTCCAACCTCCCCCCCCAACGAAAAATTCTGGTGCTGCAGGTGCACAAACGCAATTGCACTCGATTGCTGGAAATGCC
>RAL-136_Chr2L:5020221_5045940
AAATAAAATCCAATTCCAANNNNNNCCCCCAACGGAAAATTCTGGTGCTGCAGGTGCACAAACGCAATTGCACTCGATTGCTGGAAATGCC
>RAL-138_Chr2L:5020221_5045940
AAATAAAATCCAATTCCAACCTCCCCCCCCAACGGAAAATTCTGGTGCTGCAGGTGCACAAAAGCAATTGCACTCGATTGCTGGAAATGCC
>RAL-142_Chr2L:5020221_5045940
AAATAAAATCCAATTCCAACCTCCCCCCCCAACGAAAAATTCTGGTGCTGCAGGTGCACAAACGCAATTGCACTCGATTGCTGGAAATGCC
>RAL-149_Chr2L:5020221_5045940
AAATAAAATCCAATTCCAACCTCCCCCCCCAACGAAAAATTCTGGTGCTGCAGGTGCACAAACGCAATTGCACTCGATTGCTGGAAATGCC
```

Figure 2.7: Sample multiFASTA. File with the genomic sequence of five individuals.

Wiggle and bigWig file formats: text-based format used to store and display huge amounts of continuous quantitative data (Figure 2.8). It consists of one or more blocks, each containing a declaration line followed by lines defining data elements. The declaration line determines the type and specific options of the file (using space-separated key-value pairs). Wiggle data elements must be equally sized, hence the window size (i.e., span) is always fixed. There are two types of WIG files: *(a)* Fixed step: the distance between windows (step) is fixed and the data is stored in a single column of data values and *(b)* Variable step: the distance between windows is variable, so data is stored in two columns for genome positions and data values. BigWig files are Wiggle files in an indexed binary format, which improves data speed performance and storage (Kent et al. 2010).

A

```
fixedStep chrom=3L start=1 step=100 span=100
0.06
0.08
0.11
0.09
0.08
```

B

```
variableStep chrom=3L span=100
1       0.06
101     0.08
201     0.11
301     0.09
401     0.08
```

Figure 2.8: Wiggle file. **(A)** Fixed step. **(B)** Variable step. Note that both files contain the same exact information.

Box 2.1: (*Cont.*) Genome annotations file formats

Generic Feature Format version 3 (GFF3): tabulated text file format for storing genomic features (Eilbeck et al. 2005; Figure 2.9). It consists of a set of header lines for meta-data information (starting with #), followed by the annotations (one per line) with data distributed in 9 columns: (i) sequence/chromosome ID, (ii) source, (iii) feature type (defined in the Gene Ontology website), (iv) start, (v) end, (vi) score, (vii) strand, (viii) phase and, (ix) attributes. The attributes column is commonly used to specify relationships between annotations, although it can contain any information.

```
##gff-version 3
##sequence-region 2L 5029608 23011544
2L   FlyBase   gene    5029609 5037279 .  +  .  Name=Cg25C;ID=FBgn0000299
2L   FlyBase   mRNA    5029609 5037279 .  +  .  ID=FBtr0079002;Parent=FBgn0000299
2L   FlyBase   exon    5029609 5029746 .  +  .  ID=FBgn0000299-ex1;Parent=FBtr0079002
2L   FlyBase   intron  5029747 5030136 .  +  .  ID=FBgn0000299-in1;Parent=FBtr0079002
2L   FlyBase   exon    5030137 5030237 .  +  .  ID=FBgn0000299-ex2;Parent=FBtr0079002
2L   FlyBase   CDS     5030218 5030237 .  +  0  ID=FBgn0000299-cds1;Parent=FBtr0079002
2L   FlyBase   intron  5030238 5030741 .  +  .  ID=FBgn0000299-in2;Parent=FBtr0079002
2L   FlyBase   CDS     5030742 5030790 .  +  1  ID=FBgn0000299-cds2;Parent=FBtr0079002
2L   FlyBase   exon    5030742 5030790 .  +  .  ID=FBgn0000299-ex3;Parent=FBtr0079002
2L   FlyBase   intron  5030791 5030854 .  +  .  ID=FBgn0000299-in3;Parent=FBtr0079002
```

Figure 2.9: Generic Feature Format version 3 (GFF3). Gene annotations of the *D. melanogaster* reference genome.

Variant Calling Format (VCF) tabulated text file format for storing variation information (Danecek et al. 2011; Figure 2.10). It consists of a set header lines for metadata followed by data lines, each containing information about a position in the genome, distributed in 9 fixed columns: (i) sequence/chromosome ID, (ii) start, (iii) annotation ID, (iv) reference allele(s), (v) alternative allele(s), (vi) quality, (vii) filters, (viii) extra information and, (ix) format of the sample information; followed by a variable number of extra columns which correspond to specific information of each analyzed sample.

```
##fileformat=VCFv4.1
##contig=<ID=1,length=23011544>
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM      POS      ID   REF   ALT   QUAL   FILTER   INFO   FORMAT   RAL-149   RAL-335   RAL-357
2L          5020256  .    A     G     .      .        .      GT       0         1         0
2L          5020284  .    C     A     .      .        .      GT       0         0         0
2L          5020348  .    T     C     .      .        .      GT       0         0         1
2L          5020388  .    C     T     .      .        .      GT       0         1         0
2L          5020464  .    G     A     .      .        .      GT       0         1         1
2L          5020473  .    G     A     .      .        .      GT       0         0         1
2L          5020477  .    C     G,T   .      .        .      GT       1         0         2
```

Figure 2.10: Variant Calling format (VCF). Sample file with SNP data of three individuals.

2.4 iMKT: an R package for the Integrative McDonald and Kreitman Test

iMKT, which stands for Integrative McDonald and Kreitman Test, is an R package to compute the McDonald and Kreitman test (MKT, McDonald & Kreitman 1991) on polymorphism and divergence genomic data which can either be provided by the user or automatically downloaded from PopFly (Hervas et al. 2017) or PopHuman (Casillas et al. 2018).

The package was build and depends on R version ≥ 3.3 (R Core Team 2015). It requires ggplot2 (Chang 2012, Wickham 2016) complete library for the graphical display of results, and imports certain functions for performing the analyses from *utils* and *stats* default libraries, *knitr* (Xie 2012) package, and the following CRAN packages (`https://cran.r-project.org/`): *cowplot*, *reshape2*, *nls2*, *MASS* and *ggthemes*.

2.4.1 Implementation of four McDonald and Kreitman derived tests

iMKT includes the standard MKT (McDonald & Kreitman 1991) along with four MKT-derived methods which allow inferring the rate of adaptive evolution (α): FWW correction (Fay et al. 2001), DGRP correction (Mackay et al. 2012), asymptotic MKT (Messer & Petrov 2013, Haller & Messer 2017) and a new test named integrative MKT.

Besides the diverse α estimates, iMKT package also allows estimating 3 metrics mainly based on divergence levels. The first is the K_A/K_S ratio, also known as ω (Li et al. 1985, Nei & Gojobori 1986, Yang & Bielawski 2000), which compares the rates of non-synonymous and neutral divergence (Equation 2.3).

$$\omega = K_A/K_S \tag{2.3}$$

Then, it is possible to estimate the fraction of non-synonymous fixed differences which are truly adaptive (ω_A) using both ω and the rate of adaptive evolution (α) computed by any MKT Castellano et al. 2016, James et al. 2016; Equation 2.4):

$$\omega_A = \omega \times \alpha \tag{2.4}$$

Finally, the fraction of non-synonymous divergent sites which are in fact deleterious can be estimated from the previous expression (Equation 2.5).

$$\omega_D = \omega - \omega_A \tag{2.5}$$

Standard MKT

Under strict neutrality, the ratio of the number of selected and neutral polymorphic sites (P_i/P_0) is equal to the ratio of the number of selected and neutral divergence sites (D_i/D_0). The null hypothesis of neutrality is rejected in a MKT when $D_i/D_0 \neq P_i/P_0$. The excess of divergence relative to polymorphism for class i, is interpreted as adaptive selection for a subset of sites i. The fraction of adaptive fixations (α) is estimated as in Equation 2.6. The significance of the test can be assessed with a Fisher exact test (Fisher 1922).

$$\alpha = 1 - \frac{D_0}{D_i}\,\frac{P_i}{P_0} \tag{2.6}$$

FWW correction

The FWW correction (Fay et al. 2001) is an extension of the MKT in which low frequency polymorphisms are removed from the analysis, based on an arbitrary cut-off. In this case, α is estimated using the standard MKT equation, but considering only those polymorphic sites (for both neutral and selected classes) with a frequency above the established cutoff (x), as shown in Equation 2.7.

$$\alpha = 1 - \frac{D_0}{D_i}\,\frac{P_{i>x}}{P_{0>x}} \tag{2.7}$$

DGRP correction

In the DGRP correction (Mackay et al. 2012) P_i, the count off segregating sites in class i, is separated into the number of neutral variants and the number of weakly deleterious variants: $P_i = P_{i\ neutral} + P_{i\ weakdel}$.

This is done based on an established frequency cut-off (x), normally set at 5%. Consider the pair of 2×2 contingency tables from Table 2.4: the table on the left is the standard MKT table with the theoretical counts of segregating and divergent sites for each class, while the table on the right contains the count of P_i and P_0 for the two-frequency categories below and over the established cut-off.

The estimate of the fraction of sites segregating neutrally within the $frequency < x$ ($f_{neutral<x}$) is $f_{neutral<x} = P_{0<x}/P_0$. The expected number of segregating sites in the non-synonymous class which are neutral within the $frequency < x$ is $P_{neutral<x} = P_i \times f_{neutral<x}$. The expected number of neutral segregating sites in the non-synonymous class is $P_{i\ neutral} = P_{neutral<x} + P_{i>x}$. Finally, the fraction of adaptive evolution, α is estimated as in Equation 2.8:

$$\alpha = 1 - \frac{D_0}{D_i}\frac{P_{i\ neutral}}{P_0} \tag{2.8}$$

The excess of sites segregating with $frequency < x$ with respect to the neutral site class are considered to be weakly deleterious and therefore, b can be estimated as in Equation 2.9:

$$b = \frac{P_{i\ weak\ del}}{P_0}\frac{m_0}{m_i} \tag{2.9}$$

Table 2.4: Standard and DGRP MKT tables.

Standard MKT table			**Number of segregating sites by frequency category**		
Site class	Polymorphism	Divergence	Site class	$P < x$	$P \geq x$
Neutral	P_0	D_0	Neutral	$P_{0<x}$	$P_{0>x}$
Selected	P_i	D_i	Selected	$P_{i<x}$	$P_{i>x}$

Then, the neutral fraction estimated from the neutral class after correcting for weakly deleterious sites is shown in Equation 2.10:

$$f = \frac{P_{i\ neutral}}{P_0} \frac{m_0}{m_i} \tag{2.10}$$

Finally, the fraction of new mutants which are strongly deleterious and therefore not segregating is estimated as in Equation 2.11:

$$d = 1 - f - b \tag{2.11}$$

Asymptotic MKT

The asymptotic MKT method (Messer & Petrov 2013, Haller & Messer 2017) first estimates α for each derived allele frequency (DAF) category using its specific P_i and P_0 values (Equation 2.12).

$$\alpha = 1 - \frac{D_0}{D_i} \frac{P_{i(x)}}{P_{0(x)}} \tag{2.12}$$

Then an exponential function if fitted to this values (Equation 2.13:

$$\alpha_{fit}(x) = a + b^{(-cx)} \tag{2.13}$$

Although the exponential function is generally expected to provide the best fit, a linear function is also fit to the data (Equation 2.14:

$$\alpha_{fit}(x) = a + bx \tag{2.14}$$

Finally, the asymptotic α estimate is obtained by extrapolating the value of this function to 1, as shown in Equation 2.15.

$$\alpha_{asym} = \alpha_{fit}(x = 1) \tag{2.15}$$

The exponential fit is always reported, except if the exponential fit fails to converge or if the linear fit is superior according to AIC. The code of this function is adapted from Haller & Messer (2017), `http://github.com/MesserLab/asymptoticMK`.

iMKT approach

The iMKT approach is an extension of the asymptotic MKT that incorporates the estimators of the DGRP correction method to estimate diverse negative selection regimes. iMKT estimates α using the asymptotic MKT, but it only fits the exponential model (Equation 2.13) so it allows calculating the diverse fractions of mutations under purifying selection. This method provides more accurate estimates of α, although it requires better quality data to perform.

The fraction of strongly deleterious mutations (d) is estimated as the difference between neutral (0) and selected (i) polymorphic sites relative to the number of analyzed sites (Equation 2.16).

$$d = 1 - \frac{P_i/m_i}{P_0/m_0} \tag{2.16}$$

The fraction of weakly deleterious mutations (b) corresponds to the relative proportion of selected polymorphic sites that cause the underestimation of α at low DAF categories. In detail, if $\alpha_{(x)}$ is lower than the low CI estimate of α_{asym} model fitting, we considered the presence of slightly deleterious mutations in $DAF_{(x)}$ category. The weakly deleterious fraction among the segregating sites is (Equation 2.17):

$$wd = \frac{\alpha_{asym} - (\alpha_{(x)} \times (Pi_x / \sum P_i))}{\alpha_{asymptotic} - \min \alpha_{(x)}} \tag{2.17}$$

Then, the proportion of weakly deleterious mutations is estimated as in Equation 2.18:

$$b = \frac{wd}{(P_0/m_0)/(P_i/m_i)} \tag{2.18}$$

Finally, the fraction of neutral sites (f) is (Equation 2.19):

$$f = 1 - d - b \tag{2.19}$$

2.4.2 Input custom data

The initial data required to perform any MKT consists of two tables (Figure 2.11). The first one includes the number of polymorphic sites (P) in each derived allele frequency (DAF) category for both neutral (0) and putatively selected (i) functional classes. At least 10 DAF categories must be provided, but there is not any upper limit. The second table contains the number of divergent (D) and analyzed (m) sites for each class of sites. The format of input custom data is adapted from the one used by Haller & Messer (2017).

A

daf	Pi	P0
0.025	22490	17189
0.075	3217	4780
0.125	1616	2874
0.175	999	2088
0.225	754	1685
0.275	679	1443
0.325	575	1264
0.375	484	1232
0.425	427	1148
0.475	437	1068
0.525	378	986
0.575	341	928
0.625	310	893
0.675	335	928
0.725	315	945
0.775	297	822
0.825	326	885
0.875	369	953
0.925	448	1086
0.975	1019	1904

B

mi	Di	m0	D0
2598805	54641	620019	52537

A - DAF input sample file

Tab-delimited file with named columns for the derived allele frequency (daf), and the total number of non-synonymous and synonymous polymorphic sites (P_i and P_0 respectively).

B - Divergence input sample file

Tab-delimited file with named columns for the total number of non-synonymous analyzed and divergent sites (m_i and D_i) and the total number of synonymous analyzed and divergent sites (m_0 and D_0).

Figure 2.11: Input data for iMKT. Information corresponding to the complete chromosome arm 2R of a North American population of *D. melanogaster* and *D. simulans* as outgroup species. (Retrieved from `http:imkt.uab.cat`)

2.4.3 Retrieval and analysis of PopFly and PopHuman data

Gene-based metrics from PopFly (Hervas et al. 2017) and PopHuman (Casillas et al. 2018) genome browsers are downloaded using the *read.table* function from the *utils* R package. In detail, two data objects can be retrieved, with information of:

- *D. melanogaster*: 13,753 protein coding genes for 16 wild-derived populations and 4 meta-populations using *D. simulans* as outgroup species.
- *Homo sapiens*: 20,661 protein coding genes for the 26 populations of the 1000GP (Consortium et al. 2015) using *Pan troglodytes* as outgroup.

Then, data is processed and transformed in order to serve as input for the MKT-analysis functions.

3. Results

3. Results

3.1 PopFly: the *Drosophila* population genomics browser

In the last years, the continuous improvement of new sequencing technologies together with the decrease of the sequencing cost allowed us to obtain an unprecedented amount of genetic information, which requires the availability of optimal bioinformatics tools in order to be properly analyzed. Indeed, nowadays, the major bottleneck of large genomics studies is not anymore in the data acquisition step but in the computational handling and analysis of such data.

In this regard, a main bioinformatics issue when analyzing huge amounts of genomic data is how to visualize such information in an easy and intuitive manner from a population genomics perspective. After all, a visual display of the estimated metrics describing genome-wide variation and selection patterns is a key resource that would allow gaining a global view and understanding of the evolutionary forces shaping genome variation. Genome browsers provide an unique solution to this problem as they allow the graphical retrieval of the database content (Wang et al. 2013).

Taking advantage of the recently published *Drosophila Genome Neuxs* (DGN) project data, which includes the euchromatic genome sequences of more than one thousand *D. melanogaster* individuals, together with the enhancement of current custom genome browser frameworks, we developed a new genome browser named PopFly. This new genome browser is based on a similar instance previously developed by our group that hosts population genomics statistics for one single *D. melanogaster* population from Raleigh, North America (Ràmia et al. 2011). This browser, called PopDrowser, was build up as part of the Drosophila Genetic Reference Panel (DGRP) project and rapidly became a reference tool in the field. However, it has become outdated in terms of performance and data storage, which motivated the development of PopFly. The advantages of PopFly over PopDrowser are deeply discussed later in the work (*see 4.1.1. Population genomics browsers*).

PopFly can be defined as a population genomics-oriented genome browser that contains a complete inventory of population genetics parameters estimated from the DGN project data (Hervas et al. 2017). The browser server is designed for the automatic analysis and display of genetic variation data within and between populations along the *D. melanogaster* genome. The user-friendly graphical web interface of PopFly allows the visualization and retrieval of functional annotations, estimates of nucleotide diversity and divergence, linkage disequilibrium statistics, recombination rate metrics, a battery of neutrality tests, and population differentiation parameters, at different window sizes through the euchromatic chromosomes (chromosome arms 2L, 2R, 3L, 3R and X chromosome). Furthermore, the automated nature of the data processing pipeline (*see Materials and Methods for details*) makes this platform highly scalable, allowing the continuous updating of the database by the addition of the increasing number of new genome sequences available for this and related species.

PopFly, the *Drosophila* population genomics browser, is open and freely available at site `https://popfly.uab.cat`. In addition, since FlyBase release FB2017-04 (August 2017), it is also possible to access PopFly directly from the FlyBase webpage (`https://flybase.org`). When searching for any specific gene annotation in FlyBase, in the "Genomic location" section of the annotation report there is a link named "PopFly Genome Browser" which re-directs the user to a PopFly browser instance showing the annotation of interest together with nucleotide variation metrics of *Drosophila* meta-populations (Figure 3.1).

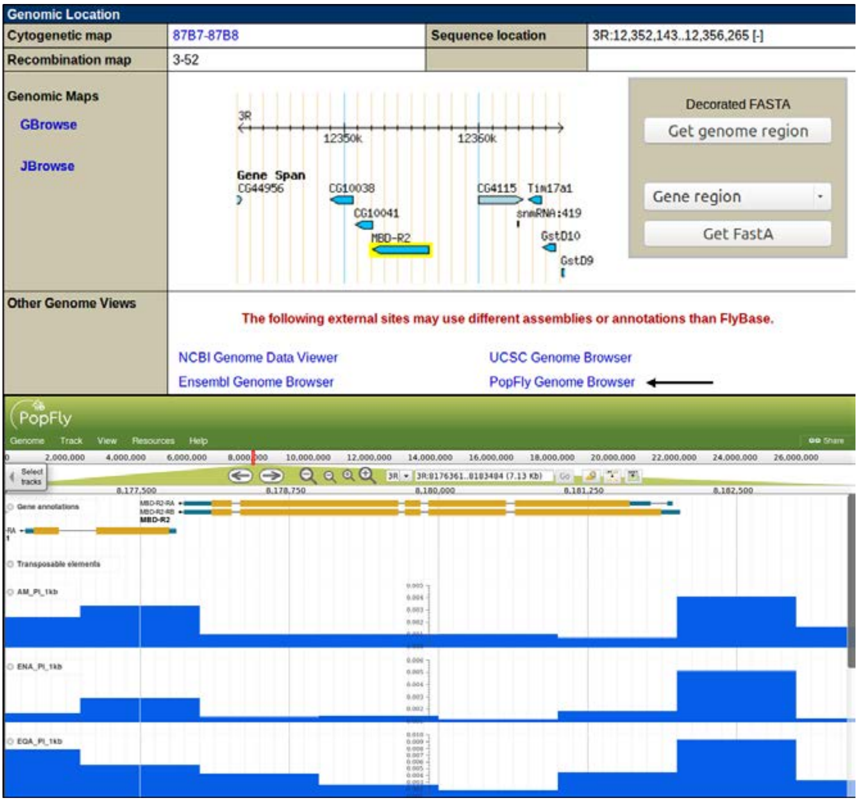


Figure 3.1: Accessing PopFly from FlyBase. The top image shows the FlyBase web report regarding gene annotation MBD-R2 which includes a link to access PopFly, located in the "Other Genome Views" section within the "Genome location" information category. Note that the section also includes links to NCBI, UCSC and Ensmbl genome browsers. The bottom image shows the PopFly browser instance accessed when clicking the corresponding link in FlyBase, with the browser view centered in the annotation of interest and the following tracks activated: gene annotations, transposable elements and nucleotide diversity metrics (π) in 10 kb windows estimated in the 6 analyzed *D. melanogaster* meta-populations. (*AUS: Oceania, CHB: China, AM: America, ENA: Europe/North Africa, EQA: Equatorial Africa, SA: Southern Africa*)

3.1.1 Browser interface

Genomic annotations and the estimated population genetic parameters are displayed along the *D. melanogaster* euchromatic chromosomes through a graphical user interface implemented in JBrowse framework (Skinner et al. 2009, Buels et al. 2016). Table 3.1 includes a summary of the annotations and main parameters estimates included in PopFly.

When accessing PopFly for the first time, a dialog box appears. It welcomes the user to the database and includes basic information about the site: where to locate instructions on how to use the server, detailed descriptions of the source data and tracks (i.e., layers of information), a Tutorial section useful for novices in population genetics or the PopFly database, contact information and the browser

Table 3.1: PopFly category tracks. Summary of tracks included in PopFly classified in 7 major categories.

Category	Annotations and main parameter estimates
Reference tracks	*D. melanogaster* reference genome (build 5.57) sequence and annotations (genes, transposable elements, polymorphic inversions, coding proportion)
Frequency-based nucleotide variation	Watterson's nucleotide diversity (θ), nucleotide diversity (π), number of 0-fold and 4-fold segregating sites (P_{0f}, P_{4f}), 0-fold and 4-fold nucleotide diversity (π_{0f}, π_{S4f})
Divergence-based metrics	Nucleotide divergence per bp (K) with *D. yakuba* and *D. simulans*, number of 0-fold and 4-fold divergent sites (D_{0f}, D_{4f}), 0-fold and 4-fold divergence (K_{0f}, K_{4f})
Linkage disequilibrium	LD sites, D, $\|D\|$, D', $\|D'\|$, r^2, number of haplotypes (h), haplotype diversity (Hd)
Recombination rate	Recombination rate estimates from Comeron et al. (2012), Fiston-Lavier et al. (2010), historical population-scaled recombination rate ($\rho_A = 2Ner$; $\rho_X = {}^8\!/\!_3 Ner$)
Selection tests based on SFS and/or variability	Fu & Li D and F test statistics, Tajima's D, Fu's Fs statistic
Selection tests based on polymorphism and divergence	K_A/K_S ratio, neutrality index (NI), direction of selection (DoS), proportion of adaptive substitutions (α) from McDonald-Kreitman test
Population differentiation	F_{ST} estimates between populations

research paper reference citation (Hervas et al. 2017).

The default browser instance (Figure 3.2) contains the navigation bar at the header of the page and the tracks panel (just below) with 5 activated tracks: (i) the reference genome sequence, (ii) gene annotations, (iii) polymorphic inversions annotations, (iv) recombination rate estimates of Comeron et al. (2012) and, (v) recombination rate estimates of (Fiston-Lavier et al. 2010). The default region displayed corresponds to the complete euchromatic chromosome arm 2L (*2L:1..23011544*).

The navigation bar includes the Help and Resources sections (with the new subsections that we have developed and implemented) along with the default JBrowse built-in modules (Genome, Track and View) and the genome coordinates box. Interestingly, at the right top area of the site there is a function ("Share") that provides the URL of the current browser instance with the actual activated tracks, which can be shared and allows restoring that specific session at any time. By default, PopFly loads the last instance viewed by the user (i.e., client-side loading), therefore it is important to supply the possibility to save any particular instance in a way it can be easily accessed over time. Finally, at the top left area of the tracks panel (and within the Resources menu) there is the tool which allows filtering and selecting tracks.

The *D. melanogaster* reference genome sequence is graphically displayed only when the view is focused on small genomic regions (< 1 kb). It contains the sequence itself for both DNA strands marked with a color code (A in green, C in blue, G in yellow, T in red) and the amino acids which correspond to the 6 possible coding reading frames (with start codons in green and stop codons in red).

Genomic annotations (such as genes, inversions or transposable elements) are displayed as horizontal rectangles covering their corresponding genomic coordinates. When right-clicking any feature, a drop-down menu with diverse options appears. All this type of tracks share an option which shows the information associated to the feature (genomic coordinates, name, id, etc.) and allows downloading its reference DNA sequence. In addition, there are specific options for each kind of annotation. For gene annotations, the options include searching for that specific feature on both NCBI and FlyBase databases and displaying a brief report of adaptation metrics based

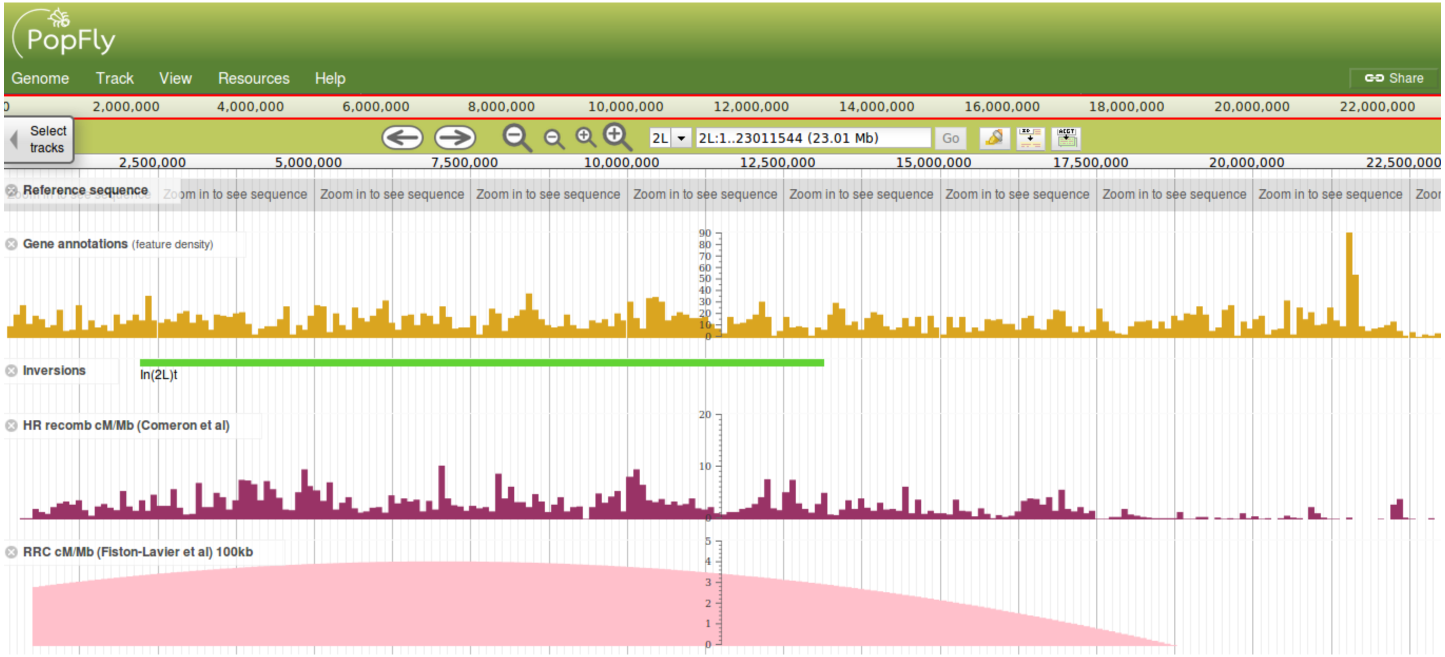


Figure 3.2: PopFly default browser instance. It contains the navigation bar at the header of the page and the tracks panel (just below) with 5 activated tracks: *(i)* the reference genome sequence (with its 6 coding reading frames, visible when zooming in), *(ii)* gene annotations, *(iii)* polymorphic inversions annotations, *(iv)* recombination rate estimates of Comeron et al. (2012) and, *(v)* recombination rate estimates of Fiston-Lavier et al. (2010). The region displayed by default is *2L:1..23011544*.

on polymorphism and divergence statistics regarding the gene of interest using user-defined populations. Moreover, gene annotations can be displayed in three modes: normal, compact and collapsed. Transposable elements annotations include an option to search for that specific feature on NCBI database. Finally, polymorphic inversions annotations include an option to display a brief report which shows their genotypes frequencies among the populations analyzed within PopFly.

Otherwise, quantitative estimates (which include most population genetic parameters) are represented as bar charts, where the height of each bar corresponds to the estimated value and the left and right limits account for the genomic coordinates of the window. These metrics are available in non-overlapping sliding windows of varying size (1 kb, 10 kb, 50 kb, 100 kb). When placing the cursor over any bar, its associated value shows up. In addition, there is also a function which permits transforming data values in a logarithmic scale, useful for certain metrics.

3.1.2 Utilities and support resources

PopFly includes all the default built-in functions of JBrowse to search and display chromosomal regions (or genes), select and filter tracks, add custom annotations, download information, and highlight specific regions of the genome. However, we have upgraded some of them to adapt their functionality to the huge amount of data stored in this server. In addition, we have also developed and implemented new utilities and support resources within the PopFly framework to facilitate performing population genetics analyses and retrieving data. Some of the upgraded and developed functions are stated below.

Help section

The Help section contains exhaustive documentation about the data analyzed by PopFly (including a brief report about the populations structure, based on genetic distances among populations using F_{ST} values) and the browser tracks. In addition, it also includes general help about JBrowse software, direct links to the welcome dialog box and the reference paper citation, and a comprehensive tutorial introducing to the usage of the PopFly database and to the testing of evolutionary hypotheses from a population genetics perspective. The tutorial contains several step-by-step guides to facilitate reproducing the results that are shown both in the form of figures and descriptive text.

Tracks selector

PopFly contains more than 4,000 tracks. Therefore, given the large number of tracks available, these can be filtered and selected using the “Select tracks” tool, which can be accessed from the top left corner, below the navigation bar or from the “Resources” section. The filtering process is performed by first narrowing the search using the menu on the left, and then selecting the tracks of interest from the main panel on the right. This process can be done several times in order to finally get the desired tracks selected. It is also possible to select a track by typing its name in the search box located at the top of the track selector tool. This utility was upgraded to handle the vast number of tracks included in the browser, classifying them in several categories and subcategories and enhancing the filtering procedure.

Retrieval of track data

Track data can be conveniently downloaded from the browser interface using the “Save track data” option which appears when pressing the arrow located to the right of the track label. This utility allows defining the region of interest (visible region or custom genomic coordinates), the output format (bedGraph, Wiggle or GFF3) and the file name. We have upgraded the initial JBrowse function to allow downloading data referring to the whole chromosome at once, as by default it is limited to genomic regions spanning a maximum of 500 kb.

Gene-based adaptation metrics report

A brief summary of adaptation metrics calculated for each protein-coding gene annotated in the reference genome can be accessed when right-clicking the feature annotation. In detail, these metrics include the K_A/K_S ratio (Li et al. 1985, Nei & Gojobori 1986), the proportion of adaptive substitutions (α) from MKT (McDonald & Kreitman 1991) along with its associated p-value from Fisher Exact test (Fisher 1922), and the direction of selection (DoS) statistics (Stoletzki & Eyre-Walker 2011). First, the population(s) of interest are selected, and then, by clicking the "Submit" button a report is generated. This brief report includes graphical display of the derived allele frequencies for neutral (4-fold degenerate) and putatively selected (0-fold degenerate) sites and a table with the number of analyzed, polymorphic and divergent sites for both functional classes, together with the metrics stated above using both *D. simulans* and *D. yakuba* outgroup species, for the selected annotation and population(s).

On-the-fly statistical analyses

The On-the-fly statistical analyses tool allows generating custom interactive plots to explore correlations among different genomic and geographical features estimated for each population or retrieved from Lack et al. (2016) (Figure 3.3A). In addition to the graphical representation of samples values, the utility provides two correlation parameters with their associated p-values: Pearson correlation coefficient (Pearson 1895) and Spearman's rank correlation coefficient (Spearman 1904). When the user clicks on a specific population, its value is removed from the graph and the statistics are re-calculated. Multiple points can be removed at once by selecting the corresponding area in the graph and clicking the "Toggle points" button. Clicking the "Reset point" button resets all points. This tool also allows deciding whether to analyze the full genome, only the autosomes or a single chromosome arm, and which parameters to correlate, from the drop-down menus located above the plot area. It can be accessed from the Resources menu placed at the navigation bar.

Download sequences

The download sequences plugin allows the easy and fast retrieval of the genome sequences (in either FASTA or VCF file formats) corresponding to any region and population(s) of interest (Figure 3.3B). The tool requires setting the genomic coordinates (chromosome, start, end) of the target region, as well as the desired population(s), and allows deciding whether or not the reference genome sequence is included in the alignment, and whether the output sequences are going to be retrieved in one single alignment file with all selected populations, or in separate files for each population. Finally, the plugin returns a zip compressed file containing the requested sequences. This tool can be accessed from the Resources section of the navigation bar or by clicking the corresponding plugin icon placed next to the genome coordinates box (the most right one).

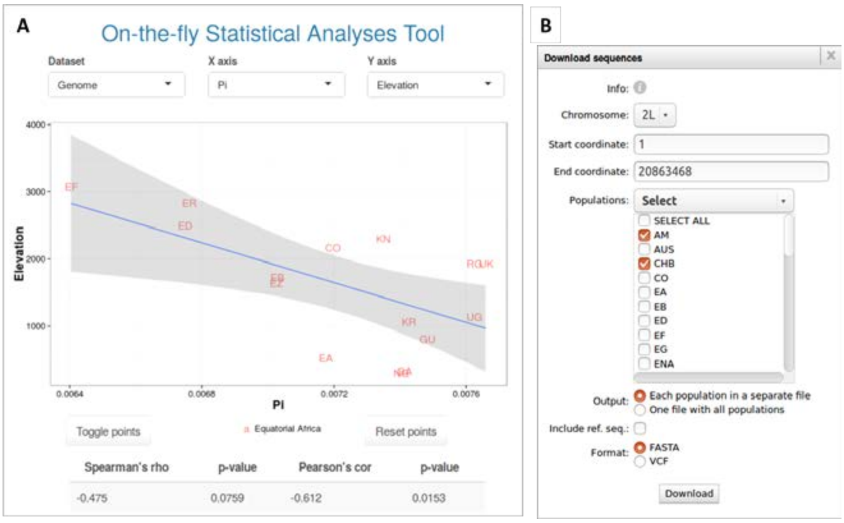


Figure 3.3: New utilities developed and implemented in the PopFly framework. (A) On-the-fly statistical analyses tool showing the correlation estimates between Elevation (corresponding to the geographical origin of samples, in meters) and nucleotide diversity (π) metrics for certain populations, considering the complete euchromatic genome. The selected populations are all from the African tropical area in order to remove the possible bias caused by differences in latitude, which are known to affect nucleotide diversity rates. In this case both correlation coefficients pinpoint a negative association between the features (Spearman's $\rho = -0.475$ with an associated $p - value$ of 0.0759 and Pearson's $r = -0.612$ with $p - value = 0.0153$, respectively). It means that higher the geographical location where the population lives, lower the polymorphism rate, probably due to harder environmental conditions and higher pressures. **(B)** The download sequences plugin with certain region (*2L:1..20863468*), populations (*AM: America, CHB: China*) and output file settings (each population in a separated FASTA file without including the reference sequence) selected.

3.1.3 Testing evolutionary hypotheses using PopFly

PopFly is designed to serve as a reference database of population genetics estimates in *D. melanogaster* and to help testing evolutionary hypotheses from a population genetics perspective. The browser allows not only to analyze and characterize any genomic region of interest, but also to perform genome-wide scans of selection in order to identify candidate regions under selective pressures, which should be then examined in detail.

This section includes an example on how to use PopFly to test a simple and direct evolutionary hypothesis. Specifically, here we analyze whether one gene of interest has been experiencing recurrent adaptive evolution in *D. melanogaster* or not, since the speciation time between *D. melanogaster* and *D. simulans*, and we describe the population genetics characteristics of the region where this gene is located.

The gene of interest that we selected is the *no hitter* (*nht*) protein-coding gene, located in the genomic region *2L:15,317,954..15,318,901*. It is a member of the Testis-specific TBP-associated factors (tTAFs) gene group, generally localized in the spermatocyte nucleoli and whose main function is to regulate the transcription of genes required for spermatocyte entry into meiosis (Hiller et al. 2004). Genes related to spermatogenesis have been reported to show signs of accelerated adaptive evolution in *Drosophila* (Ranz et al. 2003, Haerty et al. 2007). Therefore, here we use PopFly to test if the *nht* gene is under recurrent adaptive evolution in *D. melanogaster*, and if it supports the essential role of sexual selection in this species.

For this example, the ZI (Siavonga, Zambia) population is used. This population includes 197 individuals (large sample size), it shows high levels of diversity, and it is known to have a relatively simple demography, as it is widely accepted that *D. melanogaster* originated in this area (Pool et al. 2012, Lack et al. 2016). The latter is extremely important, as demographic processes can leave genomic footprints similar to those left by natural selection (e.g., a clear reduction of nucleotide diversity caused either by a population bottleneck or a selective sweep event), which may lead to wrong conclusions.

Prior to analyze the gene information itself, we described the genomic variation landscape of the genomic region where this gene is located.

For doing so, an important consideration is to select the proper visualization mode for the metrics of interest. The appropriate windows size depends on the size of the target region. In general, to inspect whole-genome patterns, it is recommended to use 100 kb or 50 kb windows, whereas if the analysis is focused on a small region, 10 kb or even 1 kb windows would be preferred. This example is centered in a small genomic region (6.4 kb), so we used the smallest windows available.

In order to get a general picture of the genomic properties of the target region, we displayed the following metrics using 1 kb windows: nucleotide diversity (π), Watterson's estimator (θ), nucleotide divergence relative to *D. simulans* (k), Tajima's D (D), linkage disequilibrium r^2; and the historical recombination rate (ρ) metrics in 10 kb windows (which are the smallest windows available for these estimates). Figure 3.4 shows the PopFly instance with the view centered in the *nht* gene and these tracks activated.

We compared the values of these six estimates for the target region (window) against the median and mean values corresponding to the chromosome arm 2L and the ZI population (Table 3.2).

Briefly, we observe that π is reduced in the target region compared to the mean or median value in the corresponding chromosome and population (ratio $\pi_{target,ZI}/\pi_{chr2L,ZI}$ of 0.8789 relative to the median and 0.8515 relative to the mean), but θ is not reduced (ratio ~ 1 in both cases). This, together with negative Tajima's D values lower than the chromosome average (absolute ratio of 1.6277 and 1.6131 relative to the median and mean values, respectively) may be indicative of an excess of rare alleles segregating at low frequencies in that region. The divergence rate relative to *D. simulans* of the candidate window is ~ 1.5 times higher than the chromosome median, but very similar (ratio ~ 1) to the mean value. The difference between the median and the mean values of divergence in the chromosome is explained by the presence of high peaks of divergence near the centromeres and telomeres of *Drosophila* chromosomes, which results in a higher average (mean) value for the chromosome. Thus, comparing the observed value to the median value is more informative in this case. However, we observe that the linkage disequilibrium estimate r^2 is lower in this region than along the complete chromosome arm (ratio = 0.7468 and ratio = 0.6837 relative to the median and mean

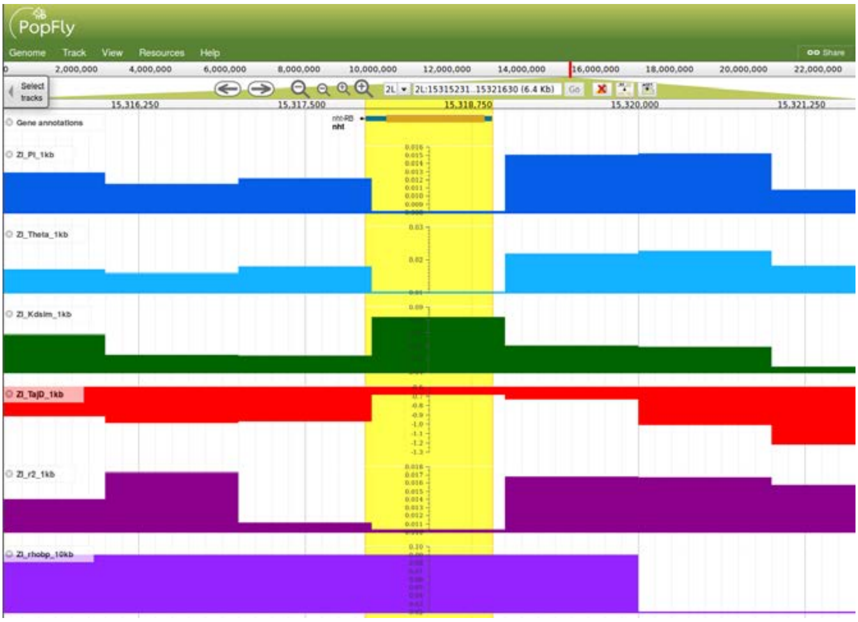


Figure 3.4: View of the gene *nht* in its genomic context using PopFly. PopFly snapshot with the view centered in the candidate genomic region *2L:15315231..15321630* and the *nhd* gene (highlighted in yellow), together with the following tracks regarding Zambia population (ZI) and 1 kb window size activated: Pi (π), Theta (θ), Kdsim (k with *D. simulans*), TajD ($Tajima's\ D$), r2 (r^2), rhobp (ρ).

values, respectively). Finally, the historical recombination rate (ρ) found in that region is extremely high (ratio = 4.2735 relative to the median and ratio = 3.4427 relative to the mean).

Then, to directly test if the *nht* gene shows signals of recurrent protein-coding adaptive evolution, we examined the adaptation metrics of this gene using the "Gene-based adaptation metrics report" utility (selecting ZI population) implemented in PopFly (Figure 3.5).

We focused on the metrics that consider *D. simulans* as outgroup species. We observe an excess of derived alleles at low frequency categories for the non-neutral class and a surplus of neutral alleles segregating at intermediate to high derived frequencies. MKT results support the hypothesis of recurrent adaptive evolution operating on the target gene, as reflected by a positive α value of 0.7719 (with an associated Fisher exact test $p = 0.0226$), direction of selection

Table 3.2: Summary metrics of the candidate region. Observed metrics values for the target region compared to the mean and median values of the chromosome arm 2L for the ZI population. Metrics estimated in 1 kb windows (except ρ, which is computed using 10 kb windows). [*D: Tajima's D estimate*]

Estimate	Target	Median	Ratio target/median	Mean	Ratio target/median
π	0.00826	0.0094	0.8789	0.0097	0.8515
θ	0.0105	0.0106	0.9905	0.0110	0.9545
k	0.082067	0.0536	1.5050	0.0759	1.0812
D	-0.6825	-0.4193	1.6277	-0.4231	1.6131
r^2	0.0104	0.0139	0.7468	0.0152	0.6837
ρ	0.0902	0.0211	4.2735	0.0262	3.4427

metric above 0 ($DoS = 0.327$) and a ratio of non-synonymous to synonymous divergence much higher than 1 ($K_A/K_S = 3.8$). Besides, we also find a signal of positive adaptation considering *D. yakuba* as outgroup species ($\alpha = 0.4118$ with $p = 0.331$; $DoS = 0.1319$ and $K_A/K_S = 1.3077$).

Summarizing, we observe that the *nht* gene is located in a region with reduced π and high levels of k (with *D. simulans*) and recombination (ρ), and an excess of rare alleles ($Tajima'sD < 0$ and $\theta > \pi$). These signatures could be indicative of an increased fixation rate of adaptive variants due to recurrent selection operating in that region. In addition, the diverse adaptation metrics estimated on the *nht* gene support the hypothesis that it is under recurrent adaptive evolution in *D. melanogaster* since the speciation time with *D. simulans*.

Hence, in this section we demonstrate that taking advantage of the huge amount of population genetics information stored in PopFly, we are able to replicate previously reported results, as well as solving new specific and concise evolutionary questions from a population genomics perspective.

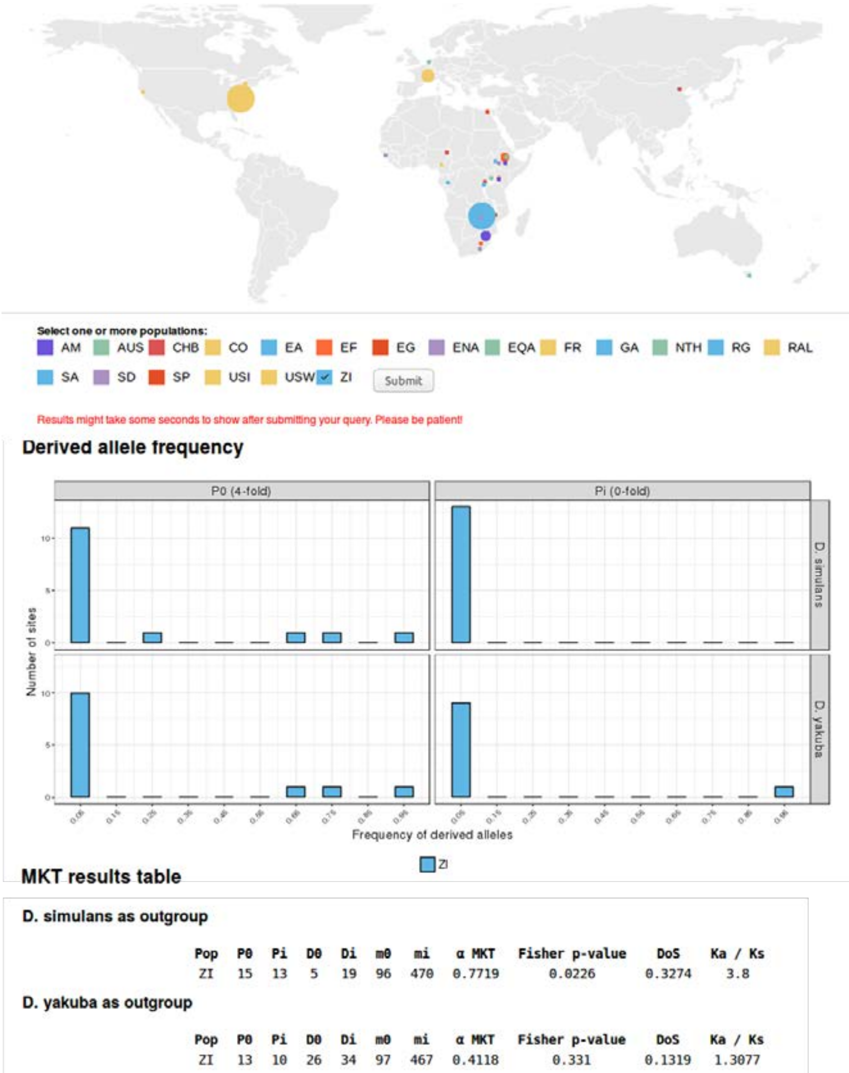


Figure 3.5: Adaptation metrics report from PopFly regarding the *nht* gene. The report shows (i) the word-wide *D. melanogaster* populations map, used to select the population of interest, (ii) a graphical display of the distribution of derived alleles (DAF) for neutral (4-fold) and putatively selected (0-fold) segregating sites and, (iii) MKT results tables using both outgroup species information. [0: neutral sites; i: putatively selected sites; P: polymorphic; D: divergent; m: analyzed]

3.2 iMKT: an R package for the Integrative McDonald and Kreitman Test

The Integrative McDonald and Kreitman Test (iMKT) package for the R software environment (a standard for statistical analyses, `https://www.r-project.org/`) allows computing the McDonald and Kreitman test (MKT, McDonald & Kreitman 1991) on polymorphism and divergence genomic data provided by the user or automatically downloaded from PopFly (Hervas et al. 2017) or PopHuman (Casillas et al. 2018). It includes five MKT derived tests which allow inferring the rate of adaptive evolution (α), as well as the fraction of strongly deleterious (d), weakly deleterious (b), and neutral (f) sites.

The package is open and freely available from GitHub at `https://github.com/BDG-UAB/iMKT` under the GNU General Public License and it can be installed using the *devtools* library (`https://github.com/r-lib/devtools`), as shown in Figure 3.6.

The iMKT package includes 12 functions (Table 3.3), classified in three main categories: (i) Calculation of MKT-derived methods; (ii) iMKT using PopFly and PopHuman data and; (iii) Miscellaneous functions. Each function has an associated help page with its corresponding description, details about input parameters, usage, examples and so on.

The first category of functions includes five different MKT derived methods (Standard MKT, FWW correction, DGRP correction,

```
## Install devtools package if necessary
> install.packages("devtools")

## Install iMKT package from GitHub
> devtools::install_github("sergihervas/iMKT")

## Load iMKT library
> library(iMKT)
```

Figure 3.6: Installation of iMKT package. The package has to be installed from GitHub using the *install_github* function from the *devtools* library. Once it is installed, it has to be loaded into the current workspace.

Table 3.3: iMKT R package functions.

Category	Function name	Description
Calculation of MKT-derived methods	standardMKT()	Original MK test (McDonald & Kreitman 1991)
	FWW()	Fay correction (Fay et al. 2001)
	DGRP()	DGRP correction (Mackay et al. 2012)
	asymptoticMKT()	Asymptotic MKT (Messer & Petrov 2013, Haller & Messer 2017)
	iMKT()	extended asymptotic MKT implementing the estimators of purifying selection from DGRP correction
	completeMKT()	Perform all previous tests at once using the same input data and parameters
iMKT using PopFly and PopHuman data	loadPopFly()	Load PopFlyData object into the current workspace, with information of 13,753 genes for 20 *D. melanogaster* populations
	loadPopHuman()	Load PopHumanData object into the current workspace, with information of 20,643 genes for 26 human populations
	PopFlyAnalysis()	Perform any MKT using PopFlyData
	PopHumanAnalysis()	Perform any MKT using PopHumanData
Miscellanea	checkInput()	Check input data and throw explicative errors when it is malformed
	themePublication()	Customization and styling of the graphs produced

asymptotic MKT, integrative MKT) which allow estimating both the rate of adaptive evolution and the fraction of sites under purifying selection; and one function to estimate all five tests at once using the same input parameters. The second one includes functions which allow the easy retrieval and analysis of population genetics information stored in PopFly (`http://popfly.uab.cat`) and PopHuman (`http://pophuman.uab.cat`) genome browsers, using the MKT methods from the previous category. Finally, the functions from the third category are used within the other functions and do not produce analyses output.

The following sections illustrate the usage of iMKT package. First we explain the new estimators from the integrative MKT method. In brief, this method is an extension of the asymptotic MKT which incorporates the estimators of the DGRP correction for quantifying negative selection. Second, we show the execution of diverse MKT-derived methods using user custom data, and third, we provide an example on how to apply iMKT using PopFly and PopHuman genomic information directly retrieved from these genome browsers.

3.2.1 Estimators of the integrative MKT

In *Drosophila*, we observe an excess of putatively selected sites relative to neutral sites segregating at low derived frequencies, while at intermediate and high frequencies the number of neutral polymorphisms is greater than the number of putatively selected ones (Figure 3.8A). This could be caused by the presence of mutations with weakly deleterious effects which have not been removed by natural selection either because they are linked to advantageous mutations or just by chance (i.e., genetic drift).

In order to provide a way to quantify with precision the fraction of such slightly deleterious mutations (b), together with the fraction of strongly deleterious (d) and neutral (f) sites within the putatively selected class, we incorporated both the asymptotic MKT and the DGRP correction methods in a unified test, named integrative MKT (iMKT). The iMKT method combines the approach developed by Messer & Petrov (2013) to estimate the fraction of adaptive substitutions (α) and an adaptation of the theoretical framework established by Mackay et al. (2012) to quantify the fraction of putatively selected sites that are under purifying selection pressures.

First, the **rate of adaptive evolution (α)** is calculated using the exponential fit of the asymptotic MKT method. The asymptotic α estimate and the α values of each derived allele frequency (DAF) category are used to quantify the fraction of weakly deleterious sites, as explained below.

The fraction of **strongly deleterious sites (d)** is estimated by comparing the rates of neutral (0) and putatively selected (i) segregating sites among the number of analyzable positions for each functional class. Assuming that all segregating sites in the selected class are effectively neutral and that there are no significant differences in the mutation rates of both classes of sites, we do not expect to observe any differences among both nucleotide variation rates. On the contrary, if we observe a larger fraction of neutral segregating sites than non-neutral ones, we can infer that this difference corresponds to the fraction of sites within the putatively selected class that are not segregating because they are strongly deleterious or even lethal (Equation 2.16).

Then, the fraction of putatively selected sites that are indeed segregating is divided into those which have weakly or slightly deleterious fitness effects and those which are effectively neutral.

The calculation of the fraction of **weakly deleterious sites (b)** is based on the observation that α estimates corresponding to low derived allele frequency categories are significantly lower than the asymptotic α value. The underestimation of α at low DAF categories is indicative of the presence of polymorphic variants with slightly deleterious effects segregating at such low frequencies. Thus, the weakly deleterious fraction among the i segregating sites is estimated as the difference between $\alpha_{asymptotic}$ and α_x corrected by the proportion of i sites segregating at that frequency among the total of segregating sites (Equations 2.17 and 2.18). In other words, b corresponds to the area between the α estimates at low DAF categories and the asymptotic α value, corrected by the number of i segregating sites in each category, i.e., the gray shaded area in Figure 3.8B.

Finally, the fraction of **neutral functional sites (f)** corresponds to the remaining proportion of segregating sites which are not slightly deleterious (Equation 2.19).

The performance (strengths, limitations, comparison with previous methods and so on) of the iMKT new estimators using empirical data from diverse populations of *D. melanogaster* is analyzed and discussed along this work.

3.2.2 Calculation of MKT-derived methods

All functions from this category require two common input parameters with polymorphism and divergence information to perform the corresponding test. The first (*daf*) has to include three named columns for the derived allele frequency (daf), the total number of non-synonymous polymorphic sites (P_i) and the total number of synonymous polymorphic sites in each frequency (P_0). The second (*divergence*) must have four named columns for the total number of non-synonymous analyzed sites (m_i), the total number of divergent non-synonymous sites (D_i), the total number of synonymous analyzed sites (m_0) and the total number of divergent synonymous sites (D_0). In addition, there are optional parameters specific for certain functions (Table 3.4).

The package includes two sample *data frames* (Figure 2.11) which can be used as *daf* and *divergence* input parameters: *myDafData* and *myDivergenceData*. Polymorphism and divergence metrics in these files correspond to the complete euchromatic chromosome arm 2R of the North American population of *D. melanogaster* from the DGRP project ($n = 205$), using *D. simulans* as outgroup species.

The output of each function always contains the corresponding α estimate, together with specific details of the selected test. This section includes two examples that illustrate how to execute and interpret the output of two functions from this category: *standardMKT* and *iMKT*. The usage of the other MKT functions is very similar to the examples presented here and can be accessed through the package documentation and each function's help page. Moreover, all five functions that perform specific MKT-derived tests have been applied to empirical genomic data and results are discussed later in this work.

Standard MKT

The first example shows how to perform a standard MKT using the *standardMKT()* function and the package sample data (Figure 3.7). The output of the function is a list which contains 4 elements:

- *Alpha estimate*: α value obtained using the original MKT (McDonald & Kreitman 1991).
- *Fisher p-value*: Fisher exact test p-value (Fisher 1922) for the MKT table, used to determine the significance of the test.
- *MKT table*: 2x2 contingency table containing the number of polymorphic and divergent sites for neutral and selected functional classes (P_0, P_i, D_0 and D_i).
- *Divergence metrics*: table with adaptation metrics based on divergence: K_a, K_s, ω, ω_A, ω_D.

As it is shown in Figure 3.7, the adaptation value of $\alpha = 0.2384$ with a corresponding Fisher's p-value of 1.4809×10^{-183} supports the hypothesis that the genomic region of interest is under adaptive evolution, with $\sim 23\%$ of fixed differences being adaptive. However, this signal can not be observed using the divergence based metrics:

Table 3.4: Input parameters for the calculation of MKT-derived methods functions from iMKT. List of parameters, with the functions on which they are required and a brief description including the default values.

Parameter	Functions	Description
daf	ALL	Data frame containing DAF, P_i and P_0 values. Mandatory. No default value
divergence	ALL	Data frame containing divergent (D) and analyzed (m) sites for selected (i) and neutral (0) classes. Mandatory. No default value
listCuttofs	DGRP(), FWW()	List of cutoffs to use for trimming. Default cutoffs values are: 0, 0.05, 0.1
plot	DGRP(), FWW(), iMKT()	Report graphical results. Options are TRUE or FALSE. Default is FALSE
xlow	asymptoticMKT(), iMKT()	Lower DAF limit for asymptotic alpha fit and estimation. Values must be in the range [0, 1]. Default value is 0
xhigh	asymptoticMKT(), iMKT()	Higher DAF limit for asymptotic alpha fit and estimation. Values must be in the range [0, 1], and *xhigh* must be higher than *xlow*. Default value is 0.9
seed	asymptoticMKT()	Seed value which allows reproducing the exact analysis. Default value is *NULL*

```
## Perform standard MKT
> standardMKT(myDafData, myDivergenceData)

> $alpha.symbol
> [1] 0.2364499

> $`Fishers exact test P-value`
> [1] 1.480943e-183

> $`MKT table`
> |               | Polymorphism| Divergence|
> |:--------------|------------:|----------:|
> |Neutral class  |        45101|      52537|
> |Selected class |        35816|      54641|

> $`Divergence metrics`
> |        Ka|        Ks|     omega|   omegaA|   omegaD|
> |---------:|---------:|---------:|--------:|--------:|
> | 0.0210254| 0.0847345| 0.2481331| 0.058671| 0.189462|
```

Figure 3.7: Execution of standardMKT() function. Output produced when executing *standardMK()* using sample *daf* and *divergence* data. It is a list of four elements: the α value, the corresponding Fisher Exact test p-value, the MKT table and five divergence-based statistics.

the ratio of non-synonymous to synonymous divergence $\omega = 0.2481$ is lower than 1, and the derived statistic $\omega_A = \omega \times \alpha = 0.0587$ is slightly different than 0.

Integrative MKT

The second example presents the execution of *iMKT()* function using again the package sample data. It uses the common *daf* and *divergence* parameters, two arguments which define the lower and higher limit for the $\alpha_{asymptotic}$ fit (*xlow*, set to 0; and *xhigh*, set to 0.9) and the option to display graphical results (*plot*) activated (*TRUE*). Haller & Messer (2017) recommend to set the higher limit for the fit at 0.9 in order to remove possible biases in the estimation of α due to polarization errors. Although it implies removing also some informative variants, the authors demonstrate that this correction provides more accurate α estimates. The output of the function (Figure 3.8) is a list that contains:

- *Asymptotic MK table*: table including information about the model type (exponential) along with the fitted function values (a, b, c), the $\alpha_{asymptotic}$ estimate with its corresponding lower and higher confidence interval values, and the $\alpha_{original}$ estimate (using the standard MKT methodology and the polymorphic sites within the xlow and xhigh cutoffs).

- *Fractions of sites*: negative selection fractions (d: strongly deleterious, f: neutral and b: weakly deleterious).

- *Graphs*: 3 plots showing: (A) the distribution of alleles frequencies (DAF) for neutral and selected sites, (B) the adaptation values (α) for each DAF category along with the function fit, the $\alpha_{asymptotic}$ and $\alpha_{original}$ estimates and the limits used for function fitting and adaptation values calculation, and (C) the negative selection fractions.

In the standard McDonald and Kreitman test, the estimate of adaptive evolution can be easily biased by the segregation of slightly deleterious non-synonymous substitutions that contribute more to polymorphism than they do to divergence and lead to an underestimation of α. The asymptotic MK method overcomes this limitation (Messer & Petrov 2013) and hence, yields to a more accurate estimation of the true level of adaptation. Therefore, in this second example (Figure 3.8), the adaptation signal appears much clearer than before ($\alpha = 0.6259$). The significance of the estimate can be assessed using the confidence intervals. As the 0 value is not included ($CI_\alpha = [0.6045, 0.6474]$), we can state that the α value is significant and that the target region (*D. melanogaster* chromosome arm 2R, in a North American population with $n = 205$ and using *D. simulans* as outgroup species) is under positive selection.

In addition, the iMKT method provides the quantification of the fraction of strongly deleterious ($d = 0.8105$), weakly deleterious ($b = 0.1271$) and neutral ($f = 0.0623$) mutations. Results show that the proportion of slightly deleterious mutations that cause the underestimation of α in the original MKT is as high as almost 13% in this specific case.

```
## Perform integrative MKT
> iMKT(myDafData, myDivergenceData, xlow=0, xhigh=0.9, plot=
    TRUE)

> $`Asymptotic MK table`
> model a b c a_asym  CI_low  CI_high a_orig
> exp 0.6259  -1.3951 18.9619 0.6259  0.6043  0.6476  0.2157

> $`Fractions of sites`
>   Type   Fraction
> 1     d 0.81053796
> 2     f 0.06232362
> 3     b 0.12713842

> $Graphs
```

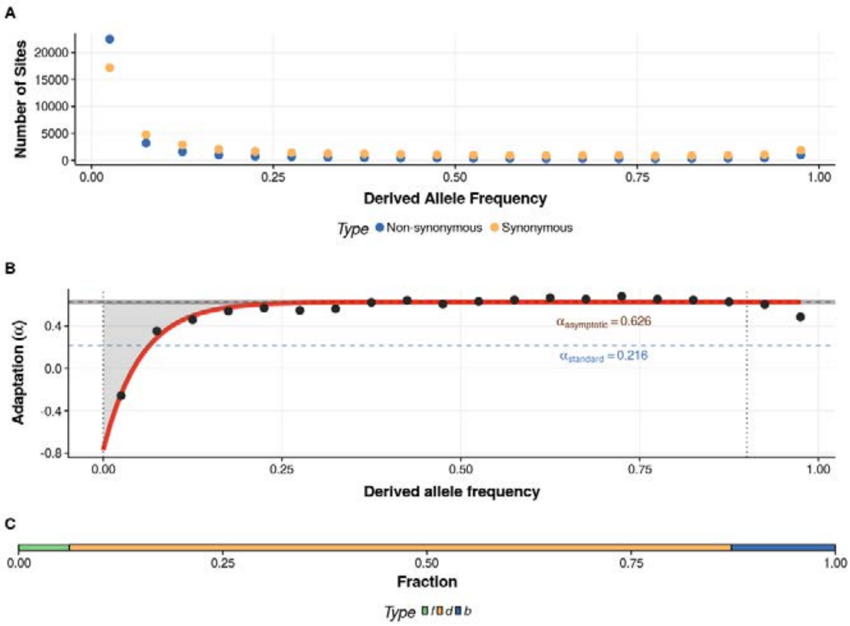


Figure 3.8: iMKT() function execution and output Execution of iMKT() using sample *daf* and *divergence* data, displaying graphical results. The output of the function includes a summary of the asymptotic model fit, with its parameters, original and asymptotic α estimates and the confidence intervals for the fit. It also contains the fractions of strongly deleterious (d), weakly deleterious (b) and neutral (f) sites within the selected class. Finally, three graphs are generated: (i) derived allele frequency counts distribution, (ii) α values and model fit and (iii) fractions of sites under purifying selection.

3.2.3 iMKT using PopFly and PopHuman data

Besides the use of custom polymorphism and divergence data, the iMKT package also allows the easy retrieval and analysis of population genetics information stored in PopFly (Hervas et al. 2017) and PopHuman (Casillas et al. 2018), the two genome browsers that contain the broadest catalog of population genetics estimates developed so far in the model species *D. melanogaster* and *Homo sapiens*, respectively. There are four individual functions devoted to do this (Table 3.3).

The *loadPopFly()* and *loadPopHuman()* functions permit the download of gene-based population genetics metrics: (i) *D. melanogaster*: 13,753 protein coding genes for 16 wild-derived populations and 4 meta-populations from the DGN (Lack et al. 2016) using *D. simulans* as outgroup and the recombination rate estimates of Comeron et al. (2012); (ii) *Homo Sapiens*: 20,661 protein coding genes for the 26 populations of the 1000GP (Consortium et al. 2015) using *Pan troglodytes* as outgroup species and recombination values retrieved from Bhérer et al. (2017) corresponding to sex-average estimates. These two functions do not have any input parameter.

The execution of each of these functions results in the loading of a new data object into the current workspace (*PopFlyData* and *PopHumanData*), which can be manually examined before starting any analysis. Each row of these *data frames* contains information regarding one gene annotation in one single population. Metrics for each gene include the number of polymorphic (P), divergent (D) and analyzed (m) positions, and the derived allele frequency distribution (DAF) for neutral (0, 4-fold) and putatively selected (i, 0-fold) sites; together with gene-associated recombination rate estimates. Once the data is loaded into the workspace, diverse iMKT analyses can be performed using the *PopFlyAnalysis()* and *PopHumanAnalysis()* functions.

These two functions allow performing any MKT using a subset of PopFly or PopHuman data defined by custom genes and populations lists. In addition, they also permit deciding whether to analyze genes grouped by recombination bins or not. They have eight input parameters, listed in Table 3.5. Briefly, each of these functions groups polymorphism and divergence values of the custom genes,

Table 3.5: Input parameters for *PopFlyAnalysis()* and *PopHumanAnalysis()* functions.

Parameter	Description
genes	List of genes to analyze. Custom genes must be listed using FlyBase IDs (FBgn...) for PopFly data and using Ensembl ID (ENSG...) for PopHuman data. Mandatory. No value by default
pops	List of populations to analyze. Available populations from PopFly are: *AM, AUS, CHB, EA, EF, EG, ENA, EQA, FR, RAL, SA, SD, SP, USI, USW, ZI*; and from PopHuman are: *ACB, ASW, BEB, CDX, CEU, CHB, CHS, CLM, ESN, FIN, GBR, GIH, GWD, IBS, ITU, JPT, KHV, LWK, MSL, MXL, PEL, PJL, PUR, STU, TSI, YRI.* Mandatory. No value by default
recomb	Group genes according to recombination values (TRUE/FALSE). Default value is FALSE. When set to TRUE, genes are sorted by recombination values and grouped in the number of bins defined by the *bins* parameter. A “concatenated” gene is created for each bin, grouping the counts of polymorphic, divergent and analyzed sites
bins	Number of recombination bins to use. Default value is 0. Mandatory only if *recomb=TRUE*
test	Which test to perform. Options include: standardMK, FWW, DGRP, asymptoticMK, iMK. Default value is standardMK
xlow	Lower DAF limit for asymptotic alpha fit and estimation. Default value is 0
xhigh	Higher limit for asymptotic alpha fit and estimation. Default value is 0.9
plot	Whether or not to report graphical results. Default value is FALSE

creating a new “concatenated gene” for each population of interest and performs the test defined, taking advantage of the functions from the previous category.

The next example shows how to analyze a subset of *D. melanogaster* genes using the *loadPopFly()* and the *PopFlyAnalysis()* functions. Following the example presented in the *3.1.3. Testing evolutionary hypotheses using PopFly* section, focused on the *nht* gene, here we focus the analysis on the gene group to which it belongs. Genes involved in sexual reproduction processes are a good target for being under adaptive evolution due to their direct implication in the evolutionary outcome of the species. In this regard, here we analyze the five genes from the Testis-specific TBP-associated factors (tTAFs) gene group: *cannonball*, *meiosis I arrest*, *no hitter*, *spermatocyte arrest* and *TBP-associated factor 30kD subunit* α*-2*. These are paralogs of the generally expressed TAF subunits of transcription factor IID (TFIID). They are predominantly localized to spermatocyte nucleoli and regulate the transcription of genes necessary for spermatocyte entry into meiosis. This gene group has been compiled by FlyBase curators using the following publications: Metcalf & Wassarman (2007), Chen et al. (2005), Hiller et al. (2004), Tora (2002), Aoyagi & Wassarman (2000).

Moreover, we use the two *D. melanogaster* populations from PopFly with largest sample sizes: RAL, Raleigh, USA, ($n = 205$) and ZI, Zambia, ($n = 197$). These populations have very different demographic histories and are subject to distinctive environmental pressures in the present. Hence, comparing both would allow assessing if any signal of adaptation is shared among all individuals of this species or, on the contrary, is specific of a certain geographic region with its associated conditions.

Here we do not consider genes' recombination context, as we are analyzing only five genes and therefore, it would mot make sense to split them in recombination bins. We use the DGRP method with the default cut-offs (0, 0.05 and 0.1) and the option to display graphical results set to *TRUE* (Figure 3.9). The function returns a list of lists with the default test output (DGRP in this case) for each population (RAL: Figure 3.10; ZI: Figure 3.11). The output of the *DGRP()* function is a list which contains five elements:

- *Results*: α estimates with their associated p-value from the Fisher Exact test for each cut-off (0, 0.05 and 0.1).
- *Divergence metrics*: list with global divergence metrics (K_A, K_S, ω) and divergence-based estimates by cut-off (ω_A, ω_D).
- *MKT tables*: 2x2 contingency tables containing the number of polymorphic sites for neutral and selected functional classes below and above each cut-off (P_0, P_i) together with the original MKT table with the number of polymorphic and divergent (D_0, D_i).
- *Fractions*: fractions of functional sites under purifying selection (d: strongly deleterious, f: neutral and b: weakly deleterious).
- *Graphs*: two graphical results: (i) α estimates for each cut-off and (ii) fractions of sites under purifying selection.

```
## List of genes
> mygenes <- c("FBgn0011569","FBgn0014342","FBgn0041103",
    "FBgn0002842","FBgn0031623")

## Execute PopFlyAnalysis
> PopFlyAnalysis(genes=mygenes, pops=c("RAL","ZI"),
    recomb=FALSE, test="DGRP", plot=TRUE)
```

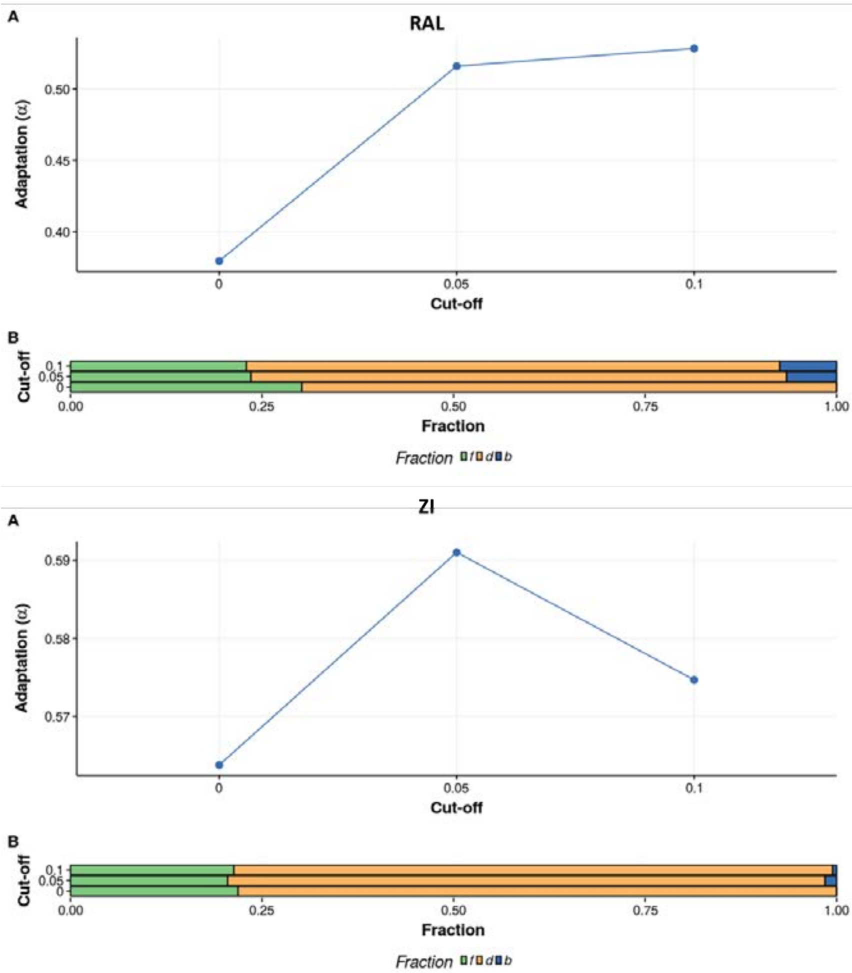


Figure 3.9: Execution of *PopFlyAnalysis()* and graphical display of results. First the genes to use are defined and stored in a list named *mygenes*. Then the *PopFlyAnalysis()* function is executed, selecting *RAL* and *ZI* populations, DGRP test, not to use recombination information and the graphical output activated. The graphs represent (A) the α estimates for each cut-off and (B) the fractions of sites under purifying selection for RAL (top) and ZI (bottom) *D. melanogaster* populations.

```
> $‘Population = RAL‘

> $‘Population = RAL‘$Results

                   alpha.symbol Fishers exact test P-value
> Cutoff =  0         0.3795094                0.081388456
> Cutoff =  0.05      0.5160173                0.007687381
> Cutoff =  0.1       0.5284271                0.007019258

> $‘Population = RAL‘$‘Divergence metrics‘

> $‘Population = RAL‘$‘Divergence metrics‘$‘Global metrics‘
          Ka         Ks     omega
> 1 0.0501912 0.1031175 0.486738

> $‘Population = RAL‘$‘Divergence metrics‘$‘Estimates by
    cutoff‘
                   omegaA.symbol omegaD.symbol
> Cutoff =  0          0.1847216     0.3020163
> Cutoff =  0.05       0.2511652     0.2355727
> Cutoff =  0.1        0.2572055     0.2295324

> $‘Population = RAL‘$‘MKT tables‘

> $‘Population = RAL‘$‘MKT tables‘$‘Number of segregating
    sites by DAF category - Cutoff = 0‘
> |               | DAF.below.cutoff| DAF.above.cutoff|
> |:--------------|----------------:|----------------:|
> |Neutral class  |                0|               33|
> |Selected class |                0|               50|

> $‘Population = RAL‘$‘MKT tables‘$‘Number of segregating
    sites by DAF category - Cutoff = 0.05‘
> |               | DAF.below.cutoff| DAF.above.cutoff|
> |:--------------|----------------:|----------------:|
> |Neutral class  |               14|               19|
> |Selected class |               32|               18|

> $‘Population = RAL‘$‘MKT tables‘$‘Number of segregating
    sites by DAF category - Cutoff = 0.1‘
> |               | DAF.below.cutoff| DAF.above.cutoff|
> |:--------------|----------------:|----------------:|
> |Neutral class  |               15|               18|
> |Selected class |               35|               15|

> $‘Population = RAL‘$‘MKT tables‘$‘MKT standard table‘
> |               | Polymorphism| Divergence|
> |:--------------|------------:|----------:|
> |Neutral class  |           33|         86|
> |Selected class |           50|        210|

> $‘Population = RAL‘$Fractions

            0       0.05        0.1
> d 0.6979837 0.69926494 0.69633630
> f 0.3020163 0.23557274 0.22953242
> b 0.0000000 0.06516231 0.07413128
```

Figure 3.10: Output of *PopFlyAnalysis()* for RAL population and DGRP correction. The output includes (i) the α and Fisher exact test p-value estimates for each cut-off, (ii) a battery of divergence-based metrics, (iii) the diverse MKT 2x2 contingency tables, and (iv) the fractions of sites under purifying selection.

```
> $`Population = ZI`

> $`Population = ZI`$Results

                alpha.symbol Fishers exact test P-value
> Cutoff =  0      0.5637718              6.819912e-05
> Cutoff =  0.05   0.5910360              2.234582e-05
> Cutoff =  0.1    0.5746775              5.921311e-05

> $`Population = ZI`$`Divergence metrics`

> $`Population = ZI`$`Divergence metrics`$`Global metrics`
          Ka          Ks     omega
> 1 0.04923518 0.09820359 0.5013583

> $`Population = ZI`$`Divergence metrics`$`Estimates by cutoff
                omegaA.symbol omegaD.symbol
> Cutoff =  0        0.2826516     0.2187066
> Cutoff =  0.05     0.2963208     0.2050375
> Cutoff =  0.1      0.2881193     0.2132390

> $`Population = ZI`$`MKT tables`

> $`Population = ZI`$`MKT tables`$`Number of segregating sites
    by DAF category - Cutoff = 0`
> |               | DAF.below.cutoff| DAF.above.cutoff|
> |:--------------|----------------:|----------------:|
> |Neutral class  |                0|               73|
> |Selected class |                0|               80|

> $`Population = ZI`$`MKT tables`$`Number of segregating sites
    by DAF category - Cutoff = 0.05`
> |               | DAF.below.cutoff| DAF.above.cutoff|
> |:--------------|----------------:|----------------:|
> |Neutral class  |               47|               26|
> |Selected class |               57|               23|

> $`Population = ZI`$`MKT tables`$`Number of segregating sites
    by DAF category - Cutoff = 0.1`
> |               | DAF.below.cutoff| DAF.above.cutoff|
> |:--------------|----------------:|----------------:|
> |Neutral class  |               53|               20|
> |Selected class |               60|               20|

> $`Population = ZI`$`MKT tables`$`MKT standard table`
> |               | Polymorphism| Divergence|
> |:--------------|------------:|----------:|
> |Neutral class  |           73|         82|
> |Selected class |           80|        206|

> $`Population = ZI`$Fractions

          0       0.05         0.1
> d 0.7812934 0.77994519 0.781518080
> f 0.2187066 0.20503746 0.213238953
> b 0.0000000 0.01501736 0.005242967
```

Figure 3.11: Output of *PopFlyAnalysis()* for ZI population and DGRP correction. The output includes (i) the α and Fisher exact test p-value estimates for each cut-off, (ii) a battery of divergence-based metrics, (iii) the diverse MKT 2x2 contingency tables, and (iv) the fractions of sites under purifying selection.

Results support the hypothesis that this set of five genes is under recurrent adaptive evolution in both populations, but adaptation metrics are slightly higher in the African population than in the American one ($\alpha_{RAL} \sim 0.52$ and $\alpha_{ZI} \sim 0.58$).

Specifically, RAL samples (Figure 3.9, 3.10) appear to be affected by a large fraction of weakly deleterious mutations ($b \sim 0.07$) which causes the underestimation of α in the Standard MKT ($\alpha = 0.3795$; $p-value = 0.081$). When slightly deleterious mutations are removed using the DGRP test, α estimates increase up to $\alpha \sim 0.52$, with an associated Fisher $p - value \sim 0.007$. The three MKT tables with the number of polymorphic sites below and above the cut-off show that 60% of sites from the selected class have a $DAF < 5\%$, and 70% of them have $DAF < 10\%$, whereas only $\sim 45\%$ of neutral sites are below this frequency cut-offs.

On the other hand, African samples from the ZI population (Figure 3.9, 3.11) do not carry such a large fraction of weakly deleterious mutations ($b \sim 0.01$) and therefore, α values are not so underestimated with the Standard MKT: $\alpha_{original} = 0.5637$; $p-value = 6.82 \times 10^{-5}$ and $\alpha_{DGRP\ 0.05} = 0.5910$; $p - value = 2.23 \times 10^{-5}$. In fact, using a cut-off of 0.1 yields to an estimate of $\alpha = 0.5747$; $p - value = 5.92 \times 10^{-5}$, closer to the one obtained with the original test. In this case, $\sim 75\%$ of putatively selected and $\sim 70\%$ of neutral sites have a $DAF < 10\%$, a smaller difference than in the previous situation.

Overall, we detect a positive selection signal in the two populations analyzed, which are known to be under very different environmental pressures. Hence, we can conclude that (i) these 5 genes as a whole have been under recurrent positive selection in *D. melanogaster* since the speciation time between this species and *D. simulans*, and that (ii) sexual selection in *Drosophila* is orthogonal to environmental variables. In addition, the fraction of slightly deleterious mutations is larger in the New World population (RAL) as expected, due to its demographic history influenced by a population bottleneck prior to a population expansion (Stephan & Li 2007). The latter may explain the observed results on which $\alpha_{RAL} < \alpha_{ZI}$.

3.3 Population genomics analyses in *Drosophila melanogaster*

PopFly stores the broadest catalog of population genetics metrics computed in *D. melanogaster* so far, with information about 30 worldwide wild-derived natural populations and 6 meta-populations covering all continents where this species inhabits (Table 2.1 and Table 2.2). Population genomics analyses over this data set would allow obtaining a global description of genetic variation patterns and assessing the major determinants of genome evolution in different populations of this model organism, which live under diverse environmental conditions and have different demographic histories. Thus, this huge population genetics comparative would enlighten our knowledge about how evolution works at the molecular level and which are the major forces causing the evolutionary change.

Figure 3.12 represents the phylogenetic tree reconstruction for DGN genome data based on F_{ST} values retrieved from Lack et al. (2016). Results show a clear division in the overall data set, with two main groups corresponding to African (ancestral) and non-African (colonizer) populations which are highly differentiated among them ($F_{ST} > 0.18$ in all pair-to-pair comparisons).

As expected, populations belonging to the same country are clustered together (South Africa: SD, SP; Ethiopia: EA, EF; United States: RAL, USI, USW). Surprisingly, the Egypt population (EG) falls within the non-African cluster, even though it is an African population. This could be due to the fact that the Sahara desert acts as a natural barrier which separates both African regions, causing higher genetic differentiation between North and South Africa populations than between North Africa and European populations. Therefore, based on our results and following Lack et al. (2016) proposal, we decided to group Egypt together with France (FR) and The Netherlands (NTH) populations in Europe/North Africa meta-population.

It is also remarkable that China population (CHB) appears to be the most differentiated population among the colonizer ones, and the fact that Australia population is genetically closer to American

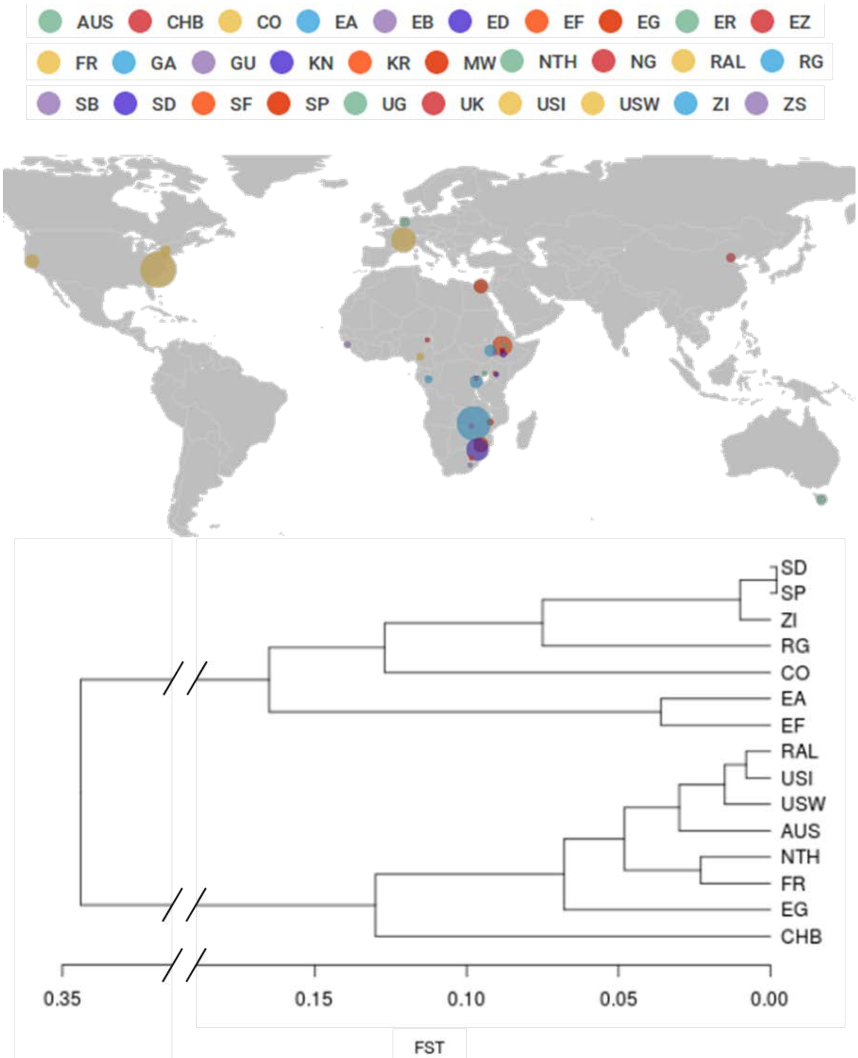


Figure 3.12: Phylogenetic tree reconstruction from F_{ST} values. F_{ST} values between pairs of 15 different populations retrieved from Lack et al. (2016). Values were averaged across chromosome arms X, 2L, 2R, 3L, and 3R, each of which was analyzed using inversion-free genomes only. *AUS: Australia, CHB: China, CO: Congo, EA: Ethiopia, EF: Ethiopia, EG: Egypt, FR: France, NTH: Netherlands, RG: Rwanda, RAL: United States, SD: South Africa, SP: South Africa, USI: United States, USW: United States, ZI: Zambia.*

than Asian populations.

Regarding the African cluster, Rwanda population (RG) seems to be closer to Southern than Equatorial Africa populations, although geographically it is more proximal to the second group. Again, geographical barriers causing the suppression of genetic exchange among populations may explain the observed unexpected genetic differences. However, we included this population into the Equatorial Africa meta-population, following Lack et al. (2016).

Finally, America and Southern Africa meta-populations composition is supported by the clustering of the corresponding populations in the reconstructed tree without any disagreement.

The following sections include global descriptions of genome-wide patterns of nucleotide variation, divergence and historical recombination, summaries of adaptation metrics estimated using diverse MKT-derived methods, and the identification and characterization of the determinants of genome variation.

In order to provide a general picture of the genetic variation landscape in *D. melanogaster* genome the population genetics analyses are focused on the 6 previously described meta-populations (Asia, Oceania, America, Europe/North Africa, Equatorial Africa and Southern Africa), using as units of analysis both 50 kb non-overlapping windows and gene-based metrics.

3.3.1 Genome-wide polymorphism and divergence patterns

For the description of genome-wide patterns of nucleotide variation we used nucleotide diversity (π) and divergence (k) with both *D. simulans* and *D. yakuba* outgroup species.

Nucleotide diversity

Polymorphism (π) values along the genome, estimated in 50 kb non-overlapping windows are shown in Figure 3.13. The distribution of π values across each chromosome arm follows a similar pattern in the diverse meta-populations analyzed.

All they show a reduction of diversity around the centromeric and telomeric regions of the autosomes. This reduction is gradual in the centromeric region, spanning several kilobases from the centromere at both arms, and more pronounced in chromosome 3 than 2 and in the left arm than in the right one. Besides, in the telomere the reduction of polymorphism is sharp and only spans few kilobases. Chromosome X behaves differentially to autosomes. In this case, there is a decrease of variability in the telomeric region which spans few kilobases and then gets stabilized, but we do not observe any reduction of diversity in the centromeric region.

Table 3.6 shows the π values estimated for each meta-population considering the complete euchromatic genome, only autosomes, X chromosome and each autosomic chromosome arm separately. Moreover, since male individuals only carry one X chromosome for every two of each autosomes (ratio 1X:2A) while females carry two of each (2X:2A), the X/A effective population size is 3/4. Hence, X polymorphism values must be corrected by a factor of 4/3 in order to eliminate the difference in effective population sizes and obtain measures comparable to autosomes.

Nucleotide diversity estimates show a decrease of variability in non-African meta-populations compared to African ones, consistent with previous studies (Mackay et al. 2012, Pool et al. 2012, Lack et al. 2015, 2016, Grenier et al. 2015). Considering full-genome, Asia has the lowest π value, followed by Oceania, America and Europe/North Africa, which have similar polymorphism values (non-significant differences, Figure 3.14). Besides, African meta-populations exhibit higher levels of nucleotide diversity. Southern Africa samples have higher levels of π than Equatorial Africa ones, pinpointing the presence of a latitudinal π cline affecting African populations. The same patterns and differences are also observed when analyzing only the autosomic chromosomes set.

When the analyses are focused on each chromosome arm independently, the differences among meta-populations remain (Figure 3.14). Results show the same tendency described above: Asia population has the lowest levels of diversity in any chromosome, African samples the highest, and the rest of meta-populations are in between.

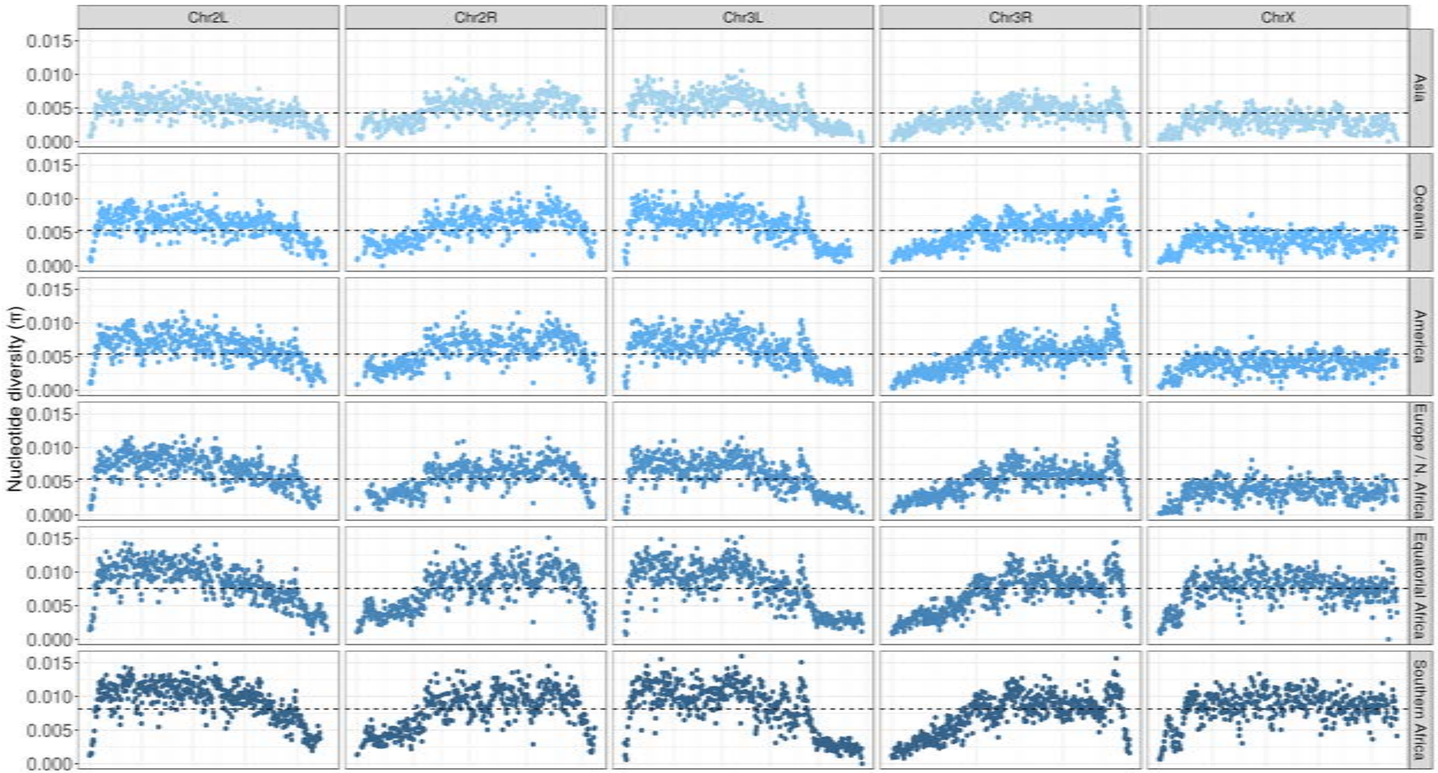


Figure 3.13: Nucleotide diversity patterns along the *D. melanogaster* genome. Polymorphism values estimated in non-overlapping windows of 50 kb covering the euchromatic X chromosome and 2L, 2R, 3L, and 3R chromosome arms. The dashed lines correspond to the mean genome π values of each meta-population.

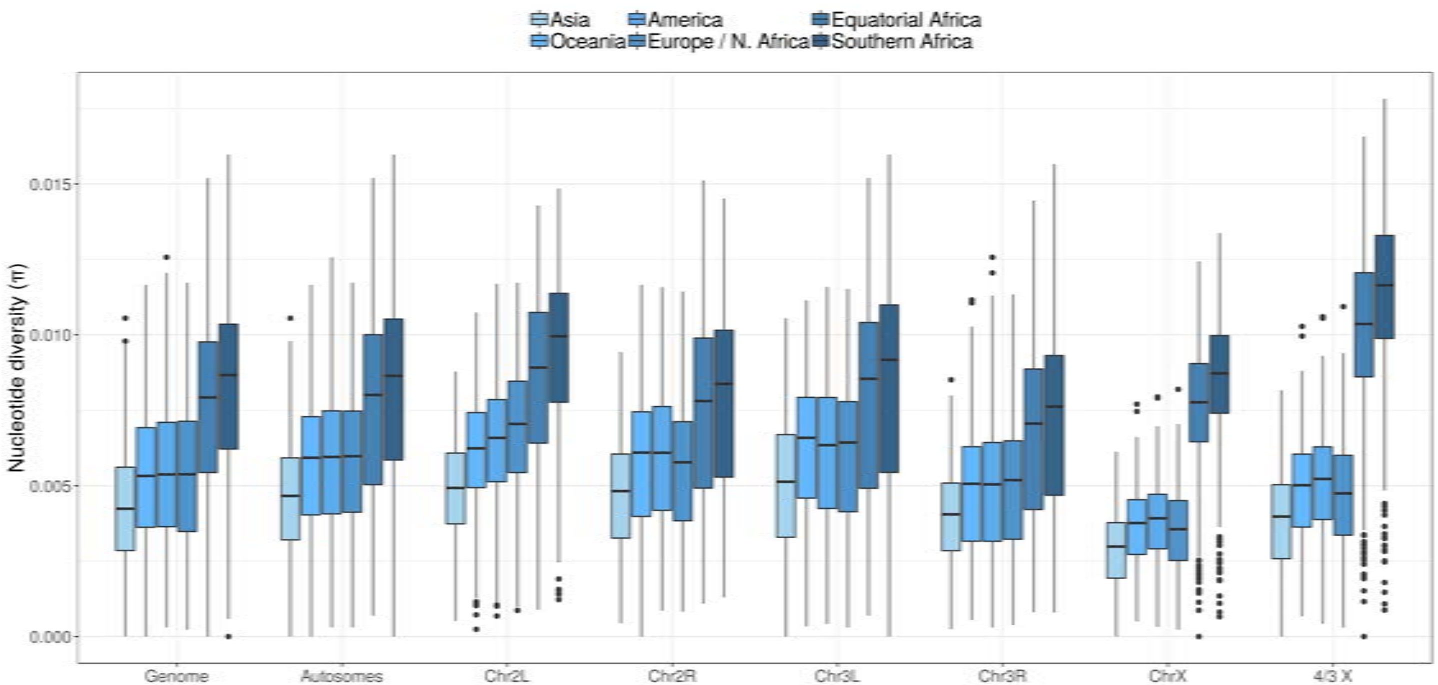


Figure 3.14: Summary and comparison of nucleotide diversity (π) estimates. (A) Box plot showing the distribution of π values for the complete genome, only autosomes, each chromosome arm and X chromosome (raw values and metrics corrected by a factor of 4/3). Values correspond to 50 kb non-overlapping windows covering the euchromatic sequence. We observe that African populations show higher values than non-African ones in all chromosomes. This tendency is much more pronounced in the X chromosome.

Table 3.6: Summary of nucleotide diversity estimates for each meta-population. Average (and standard deviation) polymorphism values (π) estimated for each meta-population regarding the complete genome (G), all autosomes (A), each chromosome arm independently, the X chromosome and X chromosome scaled by its effective population size; along with the X chromosome to autosomes ratio using raw and corrected values.

	Asia	**Oceania**	**America**	**Europe / N.Africa**	**Equatorial Africa**	**Southern Africa**
G n=2,378	0.00426 (± 0.00189)	0.00529 (± 0.00221)	0.00539 (± 0.00230)	0.00535 (± 0.00237)	0.00757 (± 0.00295)	0.00813 (± 0.00305)
A n=1,930	0.00459 (± 0.00187)	0.00568 (± 0.00219)	0.00577 (± 0000231)	0.00578 (± 0.00233)	0.00758 (± 0.00309)	0.00807 (± 0.00319)
2L n=460	0.00480 (± 0.00169)	0.00605 (± 0.00194)	0.00640 (± 0.00212)	0.00682 (± 0.00219)	0.00845 (± 0.00291)	0.00932 (± 0.00277)
2R n=422	0.00470 (± 0.00180)	0.00582 (± 0.00216)	0.00592 (± 0.00221)	0.00556 (± 0.00212)	0.00752 (± 0.00297)	0.00786 (± 0.00294)
3L n=490	0.00499 (± 0.00219)	0.00617 (± 0.00239)	0.00608 (± 0.00246)	0.00597 (± 0.00246)	0.00781 (± 0.00335)	0.00819 (± 0.00360)
3R n=558	0.00398 (± 0.00160)	0.00488 (± 0.00201)	0.00489 (± 0.00215)	0.00497 (± 0.00213)	0.00673 (± 0.00286)	0.00711 (± 0.00295)
X n=448	0.00288 (± 0.00128)	0.00362 (± 0.00132)	0.00378 (± 0.00135)	0.00347 (± 0.00143)	0.00750 (± 0.00223)	0.00840 (± 0.00233)
4X/3	0.00384 (± 0.00171)	0.00483 (± 0.00176)	0.00504 (± 0.00180)	0.00462 (± 0.00191)	0.01000 (± 0.00298)	0.01120 (± 0.00311)
X/A	0.6287	0.6368	0.6555	0.5997	0.9892	1.0411
4X/3A	0.8383	0.8491	0.874	0.7996	1.319	1.3882

We also performed pair-wise comparisons using the Wilcoxon rank sum test (Wilcoxon 1945) correcting p-values with the Holm's method (Holm 1979) to identify significant differences comparing: (i) the ratio between the X chromosome (and 4X/3) and autosomes values, for each meta-population and, (ii) meta-populations estimates for the complete genome, autosomes, X chromosome and each chromosome arm independently. The same tests were applied for divergence and historical recombination estimates (*section 3.3.2. The landscape of population historical recombination*).

The X/A ratio polymorphism estimates (Table 3.6) pinpoint a clear decrease of X chromosome variability in the non-African populations compared to autosomes (ratios $\sim$ 1.6 and $p < 10^{-16}$ in all these comparisons). However, this signal is not present in the ancestral (African) samples, which have ratios ~ 1 ($p > 0.05$ for both Equatorial and Southern Africa populations). After correcting chromosome

X values by the X effective population size (4/3), we still observe this decrease of variability in colonizer populations (ratios ~ 1.2 and $p < 10^{-10}$), whereas ancestral samples have ratios of ~ 0.7 (and $p < 10^{-16}$).

Regarding the comparisons between meta-populations, statistical analyses support the differences explained above. Asia meta-population shows lower variability than the rest, no matter the chromosome arm analyzed (all $p < 10^{-8}$). African meta-populations values are statistically higher than non-African ones (all pair-wise comparison p-values between African and non-African populations are $p < 10^{-14}$), with Southern Africa samples exhibiting more diversity than Equatorial Africa samples ($p < 10^{-5}$ for genome, autosomes, chromosome arm 2L and X chromosome; and $p > 0.05$ for 2R, 3L and 3R. Then, Oceania - America comparisons only show statistical differences in chromosome arm 2L ($p = 0.006$), but not in the rest ($p > 0.05$). Europe/N. Africa - America comparisons show a similar pattern to the one just described, with differences in 2L chromosome arm 2L ($p = 0.0052$) and chromosome X ($p = 0.0032$) but not in the rest of comparisons (all $p > 0.05$). Finally, Europe/N. Africa - Oceania only show differences in chromosome arm 2L ($p = 10^{-8}$), with all the rest of comparisons having an associated $p > 0.05$.

In summary, as global trends we observe higher levels of nucleotide diversity in African than non-African meta-populations and a clear decrease of X chromosome polymorphism in colonizer populations, which remains even after correcting by the X chromosome effective population size. In addition, we observe that Asia meta-population presents the lowest values of π, then Oceania, America and Europe/North Africa meta-populations have intermediate and similar values and finally, Equatorial Africa and Southern Africa have the highest variability estimates, with a latitudinal polymorphism cline (souther the population, higher π).

Divergence with *D. simulans* and *D. yakuba*

Divergence per base-pair (k) metrics considering both *D. simulans* and *D. yakuba* outgroup species computed in 50 kb non-overlapping windows covering the euchromatic *D. melanogaster* genome are shown in Figure 3.15.

We observe the same pattern no matter the outgroup considered nor the meta-population analyzed, with a plain distribution of values along almost all entire chromosome arms and high peaks of divergence in the centromeric regions. These peaks could be due to (i) a reduced quality of alignments in these regions producing more sequence errors, (ii) higher mutation rates in those regions or (iii) higher fixation of slightly deleterious mutations due to low recombination reducing the efficiency of selection (Ràmia 2015). This effect is much more notorious in chromosome 2 than in 3. In fact, chromosome arm 3R shows the weakest signal of increased k in the centromeric region.

We also estimated the Spearman's correlation coefficient of divergence between *D. melanogaster* and both *D. simulans* and *D. yakuba* for all meta-populations and chromosome arms (Supplementary Figure 7.1). We observe that divergence metrics computed using both outgroup species show a high positive and very significant correlation (associated $p \sim 0$ in all cases) in all meta-populations and chromosome arms analyzed, with correlation coefficients ranging from 0.79 to 0.9 for the autosomes and from 0.59 to 0.61 for the X chromosome metrics, depending on the meta-population.

Table 3.7 shows average and standard deviation nucleotide divergence estimates (k) for each meta-population and genomic region (complete genome, autosomes, each chromosome arm, X chromosome), together with the X chromosome to autosomes ratio. We observe that divergence levels do not vary between meta-populations and are slightly higher in the X chromosome than the autosomes. In addition, estimates of divergence relative to *D. yak* are ~ 2 times higher than relative to *D. sim*, but the same pattern is observed in both cases, with a similar distribution of divergence in the autosomes and higher levels in the X chromosome (Figure 3.16). America and Europe/N. Africa meta-populations show slightly lower levels of divergence than the rest, and Equatorial Africa is the one with the highest levels of divergence.

Even though the X/A ratio of divergence seems only slightly higher than 1 (k_X is ~ 1.15 higher than k_A for *D. simulans* and ~ 1.1 higher for *D. yakuba*), the observed differences between autosomes and X chromosome estimates are significant in all meta-populations and for both outgroup species ($p < 10^{-16}$).

Then, we compared divergence metrics for meta-population. Using

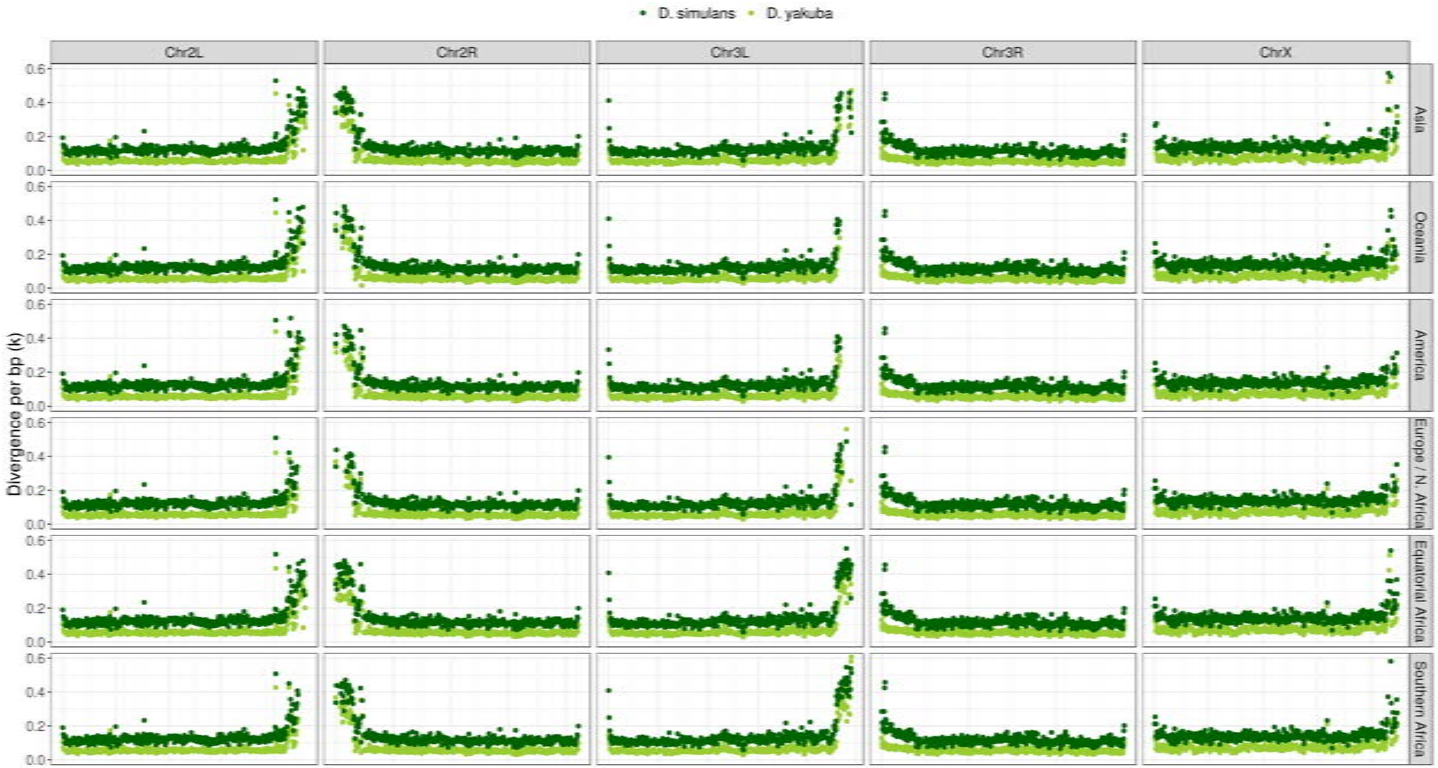


Figure 3.15: Nucleotide divergence per bp between *D. melanogaster* and *D. simulans* and *D. yakuba* along the *D. melanogaster* genome. Nucleotide divergence values estimated in non-overlapping windows of 50 kb covering the euchromatic X chromosome and 2L, 2R, 3L, and 3R chromosome arms. Light and dark green dots correspond to divergence relative to *D. simlans* and *D. yakuba*, respectively.

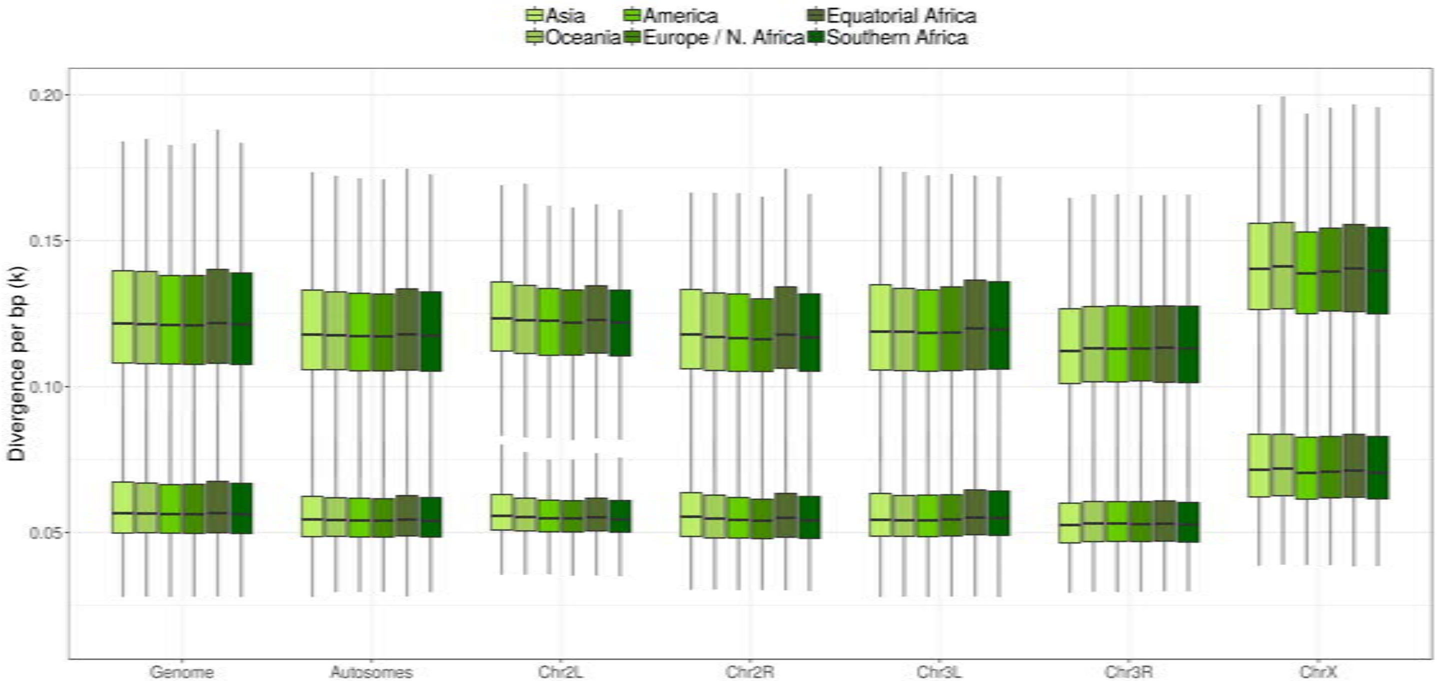


Figure 3.16: Divergence per bp (k) estimates for each chromosome arm and meta-population. Divergence estimates using *D. yakuba* (top) and *D. simulans* (bottom) outgroup species. Box plot showing the distribution of k values for the complete genome, all autosomes, each autosomic chromosome arm and X chromosome. Values correspond to 50 kb non-overlapping windows covering the euchromatic sequence. We observe a clear increase of divergence in the X chromosome relative to the autosomes, a tendency maintained in all meta-populations.

Table 3.7: Summary of divergence per bp (k) estimates for each meta-population. Average (and standard deviation) values of divergence (k) with *D. simulans* and *D. yakuba* estimated for each meta-population, along with the X chromosome to autosomes ratio. *G: Genome; A: Autosomes.*

		Asia	Oceania	America	Europe / N.Africa	Equatorial Africa	Southern Africa
D. simulans	G n=2,378	0.06743 (± 0.04713)	0.06485 (± 0.03802)	0.06396 (± 0.03678)	0.0633 (± 0.03421)	0.06982 (± 0.05399)	0.06713 (± 0.04801)
	A n=1,930	0.06492 (± 0.0486)	0.06209 (± 0.0398)	0.0616 (± 0.03915)	0.06062 (± 0.03591)	0.06752 (± 0.05504)	0.06503 (± 0.05117)
	2L n=460	0.06969 (± 0.05334)	0.06733 (± 0.049)	0.06536 (± 0.04738)	0.06099 (± 0.03133)	0.06885 (± 0.0529)	0.06227 (± 0.03633)
	2R n=422	0.07313 (± 0.06564)	0.06899 (± 0.05824)	0.06853 (± 0.05705)	0.06436 (± 0.0457)	0.0758 (± 0.06995)	0.07126 (± 0.06237)
	3L n=490	0.06514 (± 0.05029)	0.0593 (± 0.02689)	0.0597 (± 0.02879)	0.06365 (± 0.04612)	0.07323 (± 0.06788)	0.07381 (± 0.07246)
	3R n=558	0.05467 (± 0.01445)	0.0551 (± 0.01438)	0.05508 (± 0.01429)	0.05506 (± 0.01435)	0.05515 (± 0.01444)	0.05489 (± 0.01438)
	X n=448	0.07822 (± 0.03838)	0.07664 (± 0.02612)	0.07411 (± 0.02127)	0.07475 (± 0.02235)	0.07975 (± 0.04802)	0.07617 (± 0.02927)
	X/A	1.1601	1.1818	1.1587	1.1809	1.1422	1.1346
D. yakuba	G n=2,378	0.13351 (± 0.05562)	0.13121 (± 0.04872)	0.12983 (± 0.04637)	0.12893 (± 0.04345)	0.13622 (± 0.0629)	0.13332 (± 0.05713)
	A n=1,930	0.13061 (± 0.05789)	0.12795 (± 0.05097)	0.12709 (± 0.04947)	0.12582 (± 0.04586)	0.13379 (± 0.06571)	0.1308 (± 0.06051)
	2L n=460	0.1392 (± 0.06399)	0.13642 (± 0.05922)	0.13345 (± 0.05482)	0.1289 (± 0.04097)	0.13778 (± 0.06216)	0.13014 (± 0.04509)
	2R n=422	0.13853 (± 0.07429)	0.13436 (± 0.06654)	0.13359 (± 0.06578)	0.12888 (± 0.05474)	0.14184 (± 0.08062)	0.13643 (± 0.07139)
	3L n=490	0.12995 (± 0.05525)	0.12501 (± 0.04081)	0.12516 (± 0.0411)	0.12864 (± 0.05312)	0.14012 (± 0.0785)	0.14052 (± 0.08087)
	3R n=558	0.11817 (± 0.03363)	0.11881 (± 0.03357)	0.11883 (± 0.0337)	0.11876 (± 0.03357)	0.11888 (± 0.03363)	0.1186 (± 0.03361)
	X n=448	0.14602 (± 0.04237)	0.14512 (± 0.0344)	0.14158 (± 0.02647)	0.14224 (± 0.02738)	0.14672 (± 0.04764)	0.14415 (± 0.03757)
	X/A	1.0937	1.106	1.0905	1.1032	1.0771	1.0812

either *D. simulans* or *D. yakuba* as outgroup species, all pair-wise comparisons of divergence metrics had an associated $p > 0.05$. Thus, we do not observe statistically significant differences between populations, as expected.

Overall, divergence metrics do not vary between meta-populations and autosomic chromosome arms but are higher in the X chromosome than in the autosomes. We observed the same pattern using both *D. simulans* and *D. yakuba* outgroup species. However, K_{Dyak} metrics are ~ 2 times higher than K_{Dsim} because of the speciation time between *D. melanogaster* and each of these species.

3.3.2 The landscape of population historical recombination

The rate of population-scaled historical recombination per base-pair (ρ/bp) for autosomes and the X chromosome ($\rho_A = 2\ N_e\ r$ and $\rho_X = 8/3\ N_e\ r$, respectively) is represented in Figure 3.17. Metrics were estimated in 50 kb non-overlapping windows along the euchromatic genome of *D. melanogaster*.

We observe a clear decrease of ρ in all centromeric and telomeric regions of the autosomes, similar to the one observed for nucleotide diversity (π) and consistent with Chan et al. (2012) results and other recombination rate estimates such as Fiston-Lavier et al. (2010) and Comeron et al. (2012). The comparison to previous estimates of recombination in *D. melanogaster* is addressed in the next chapter (*Discussion*) of this work.

The pattern is variable between meta-populations, but there are some similar trends and regions with increased rates of recombination (e.g., in the middle area of chromosome arm 2L there is a clear peak of high ρ shared between Oceania and America meta-populations which is also observed in other populations but with a weaker signal). Surprisingly, Asia estimates show a high rate of dispersion, with extremely high and low values distributed along the genome, but the decrease of ρ in the telomeric and centromeric regions is also visible. Besides, the X chromosome shows a drastic decrease of ρ at both centromeric regions and a hilly distribution of recombination estimates along its middle region.

Table 3.8 shows average (and SD) ρ metrics estimated for each meta-population. Regarding full genome estimates, we observe that $\rho\ Asia < \rho\ Oceania < \rho\ America < \rho\ Europe/N.Africa < \rho\ EquatorialAfrica < \rho\ SouthernAfrica$. However, this order is not maintained in all chromosomes. Equatorial Africa shows higher values for each autosomic chromosome arm than Southern Africa, but X chromosome estimates in the Southern Africa meta-population are extremely higher than in the rest. In addition, there are certain situations in which the order stated above is changed, such as chromosome arm 2L where $\rho\ Asia > \rho\ Oceania$ and $\rho\ America > \rho\ Europe/N.Africa$. Again, we observe a high dispersion of Asia

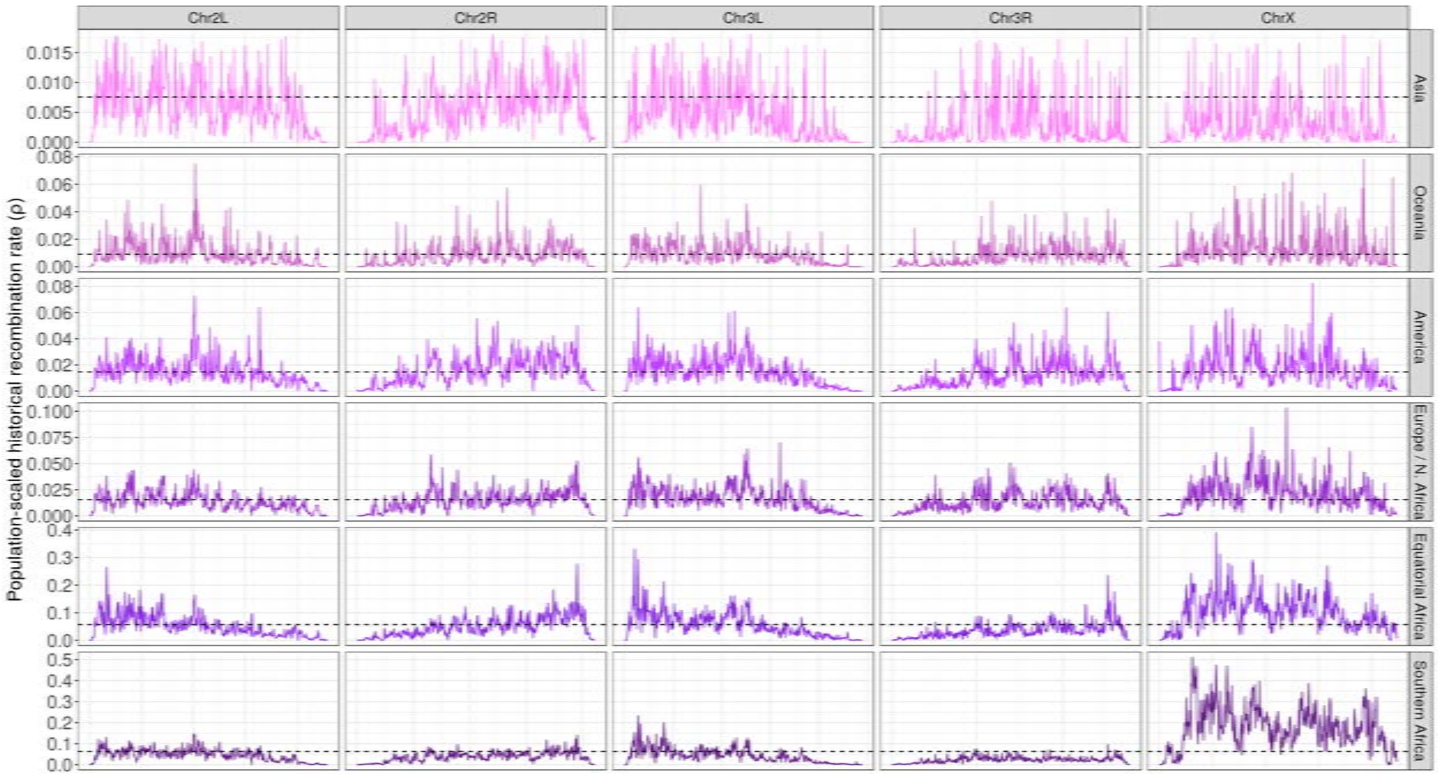


Figure 3.17: Population-scaled historical recombination rate (ρ) patterns along the *D. melanogaster* genome. Recombination values estimated in non-overlapping windows of 50 kb covering the euchromatic X chromosome and 2L, 2R, 3L, and 3R chromosome arms. The dashed lines correspond to the mean genome ρ values of each meta-population. Note that the Y-axis scale is different for each analyzed meta-population. We observe that in all populations except Asia there is an increase of ρ in the X chromosome relative to the autosomes.

Table 3.8: Summary of historical recombination rate estimates for each meta-population. Average (and standard deviation) population-scaled historical recombination rate (ρ) estimated for each meta-population, along with the X chromosome to autosomes ratio.

	Asia	**Oceania**	**America**	**Europe / N.Africa**	**Equatorial Africa**	**Southern Africa**
G n=2,378	0.00769 (± 0.01366)	0.00883 (± 0.00888)	0.01452 (± 0.01099)	0.01523 (± 0.01103)	0.05429 (± 0.04536)	0.05548 (± 0.06573)
A n=1,930	0.00796 (± 0.01396)	0.00836 (± 0.00807)	0.01425 (± 0.01056)	0.01442 (± 0.00991)	0.04747 (± 0.03821)	0.03663 (± 0.02784)
2L n=460	0.0101 (± 0.01415)	0.00933 (± 0.00928)	0.01484 (± 0.01046)	0.01361 (± 0.00871)	0.05136 (± 0.03923)	0.04519 (± 0.02862)
2R n=422	0.00688 (± 0.00687)	0.00829 (± 0.00784)	0.01549 (± 0.01096)	0.01455 (± 0.0099)	0.04981 (± 0.03736)	0.03708 (± 0.02347)
3L n=490	0.00761 (± 0.01421)	0.00857 (± 0.00769)	0.01484 (± 0.01069)	0.01639 (± 0.0115)	0.05632 (± 0.04489)	0.04278 (± 0.03576)
3R n=558	0.00733 (± 0.01708)	0.00743 (± 0.00738)	0.0123 (± 0.00997)	0.01327 (± 0.00908)	0.03473 (± 0.02674)	0.02385 (± 0.01454)
X n=448	0.0059 (± 0.01139)	0.01191 (± 0.01252)	0.01629 (± 0.01335)	0.02045 (± 0.01555)	0.09834 (± 0.06073)	0.17727 (± 0.10062)
X/A	0.7785	1.32	1.1133	1.3144	1.7236	2.8081

estimates, with SD values much higher than expected ($SD > mean$ in all chromosomes). This effect is extreme in the chromosome arm 3R ($mean = 0.00733$ and $SD = 0.01708$). Results are also graphically displayed in Figure 3.18.

The X/A ratio follows the same pattern described for the complete genome estimates, but Oceania > America. It is lower than 1 for Asia meta-population (0.7785, $p = 10^{-9}$), very similar to 1 for America ($p = 0.052$), between 1.3 and 2 for Oceania (1.32, $p = 10^{-6}$), Europe/N.Africa (1.3144, $p = 10^{-13}$) and Equatorial Africa (1.7236, $p < 10^{-16}$) and extremely high for Southern Africa (2.8081, $p < 10^{-16}$).

When analyzing ρ metrics for each chromosome arm independently (Figure 3.18), we observe the same pattern described above, with higher rates in African populations. This effect is much more pronounced in the X chromosome. In addition, we observe that African samples show higher rates of dispersion than do non-African ones. All pair-wise comparisons using either complete genome or autosomes estimates have an associate $p < 10^{-16}$ except the comparisons America (AM) - Europe/N. Africa (ENA) ($p = 0.00269$ and $p = 0.21$,

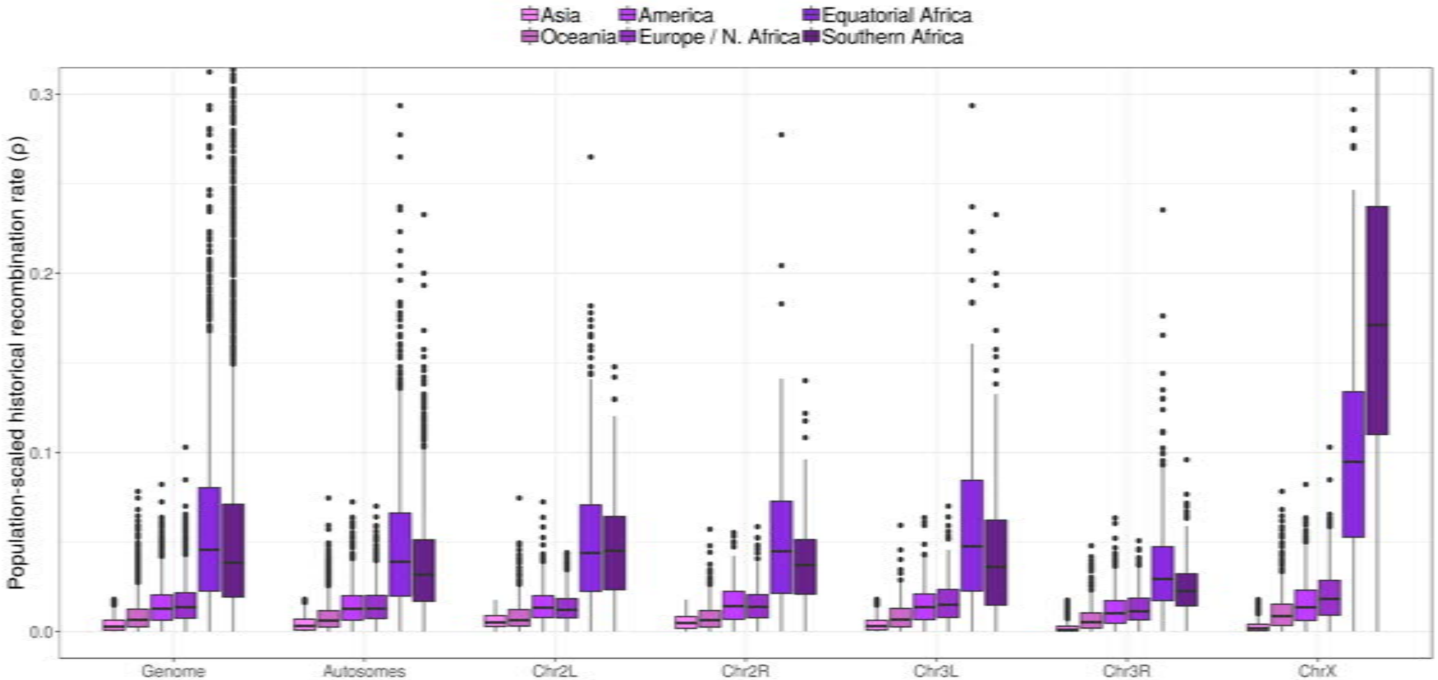


Figure 3.18: Summary and comparison of population-scaled historical recombination rate (ρ) estimates. Box plot showing the distribution of ρ values for the complete genome, autosomes, each chromosome arm and X chromosome. Values correspond to 50 kb non-overlapping windows covering the euchromatic sequence. We observe that African populations show higher values than non-African ones in all chromosomes. This tendency is much more pronounced in the X chromosome.

respectively) and Equatorial Africa (EQA) - Southern Africa (SA) ($p = 0.00034$ and $p = 10^{-15}$, respectively). For chromosome arm 2L, all comparisons have $p < 10^{-16}$ except Australia (AUS) - Asia (CHB), America - Europe/N. Africa and Equatorial Africa - Southern Africa ($p > 0.05$). The same trend is maintained in the rest of chromosomes, with all $p < 10^{-16}$ except: AUS - CHB ($p = 0.011$), AM - ENA ($p = 0.328$), EQA - SA ($p = 10^{-5}$) for chromosome arm 2R; AUS - CHB ($p = 10^{-9}$), AM - ENA ($p = 0.042$), EQA - SA ($p = 10^{-6}$) for chromosome arm 3L; AM - ENA ($p = 0.0068$), EQA - SA ($p = 10^{-12}$) for chromosome arm 3R and; AUS - AM ($p = 10^{-9}$), AM - ENA ($p = 10^{-5}$) for X chromosome.

Overall, we observe that African meta-populations show higher rates of historical recombination than do non-African ones, and that X chromosome estimates are much higher than autosomes values for these populations specifically. Non-African populations show similar ρ rates, with Asia samples presenting the lowest estimates and a X/A ratio lower than one, while the rest of populations have rates higher than 1.

3.3.3 Detection of positive and purifying selection in the *Drosophila* genome

Adaptation metrics presented in this section were estimated for 13,753 protein-coding genes annotated in the reference *D. melanogaster* genome and using *D. simulans* as outgroup species, because of the higher quality of this genome compared to the *D. yakuba* genome sequence (Hu et al. 2013). The section is divided in three parts: (i) comparative analysis of four MKT methods at the single-gene level; (ii) genome-scale estimates of adaptive evolution and; (iii) quantifying the fraction of purifying selection in the genome.

Briefly, the first and second parts show the estimates of **positive selection** both (i) at the single gene level and (ii) at the genomic scale. So first, we estimated the fraction of adaptive substitutions (α) for each single protein coding gene using four tests derived from the McDonald and Kreitman Test (MKT, McDonald & Kreitman 1991) and polymorphism and divergence genomic data from the

RAL (Raleigh) North American population ($n = 205$, the population with the largest sample size available in the *DGN* data). Second, we analyzed the adaptation rate of the six *D. melanogaster* metapopulations considering the complete set of genes as a whole and using again all four MKT methods.

Then, the third part focuses on the detection of **negative selection** at the genomic scale. We estimated the fraction of functional sites which are strongly deleterious (d), weakly deleterious (b) and effectively neutral (f) using two of these methods (DGRP correction and integrative MKT), which allowed to quantify the abundance of purifying selection in the *D. melanogaster* genome.

i. Comparative analysis of four MKT methods at the single-gene level

We estimated α for each single protein coding gene using the Standard MKT (McDonald & Kreitman 1991), FWW correction (Fay et al. 2001), DGRP correction (Mackay et al. 2012) and the Asymptotic MKT (Messer & Petrov 2013) (forcing the exponential model fit). We used non-synonymous coding sites (0-fold) as the non-neutral class (i), and synonymous coding sites (4-fold) as neutral (0).

- *Standard MKT*: in order to remove non-informative α values of 1 and $-\infty$ from the analysis, we only considered the 11,005 genes (80% of the initial set) that have at least one neutral polymorphic site (P_0) and one non-synonymous divergent site (D_i), even though a fraction of them lack variability enough to have power to be detected statistically. Results show a $\bar{\alpha} = -1.275$ (± 3.956) among the 11,005 analyzable genes. There are 739 genes with α positive and $p < 0.05$ (p determined with a Fisher exact test, Fisher 1922) with a $\bar{\alpha} = 0.789$ (± 0.138).

- *FWW correction*: we estimated α using the following cut-offs: 0.025, 0.075, 0.125, 0.175, 0.225, 0.275 and 0.325. Table 3.9 shows the number of genes available to be analyzed using each cutoff. We observe that α increases as it does the cut-off, but the number of analyzable genes decreases because we are removing polymorphic sites from the analysis, which leads to lower number of genes with enough segregating positions to be analyzed and a decrease of the test power. We obtain the highest number of genes with positive α and $p < 0.05$ using a cut-off of 0.025, with a total of 1,039 genes ($\bar{\alpha} = 0.861$; ± 0.123).

- *DGRP correction*: we used the same cut-offs as in the FWW correction. Table 3.10 shows the number of genes available to be analyzed using each cutoff. As this method does not remove polymorphic sites but only classifies them in two categories according to the established cut-off (below and over), the number of genes available to be analyzed does not decrease while increasing the cut-off. Indeed, all 11,005 genes analyzable in the Standard MKT are also analyzable with this method using any cut-off. In addition, the number of genes with positive α and $p < 0.05$ is higher than in the FWW correction. Again, we observe the highest number of positive and significant genes using a cut-off of 0.025, with a total of 1,102 genes ($\bar{\alpha} = 0.819$; ± 0.130).

- *Asymptotic MKT*: only 236 genes are analyzable with the asymptotic MKT when the exponential fitting is forced. This happens because the method requires a minimum of polymorphic sites segregating at different DAF categories in order to estimate α, and many genes do not accomplish this requirement. Out of the 236 analyzable genes, 35 are significant and positive, with a $\bar{\alpha} = 0.774$ (± 0.198). Genes are considered significant if the CI interval of the α estimate does not include the 0. Only polymorphic sites in the DAF range between 0 and 0.9 are used to avoid a bias in the α estimates, as suggested by Haller & Messer (2017).

Table 3.11 shows the α values obtained for the 236 genes analyzable by the asymptotic MKT using the other methods. Because we only used polymorphisms in the DAF range between 0 and 0.9 to estimate α with the asymptotic MKT, we analyzed this set of genes using the same polymorphism DAF range in the other tests. In this case, FWW test allows estimating positive α values with $p < 0.05$ for more than 80 genes, an increase of ~ 2.3 fold with respect to the asymptotic MKT. In addition, α estimates are higher with this method. We observe a clear increase of the number of genes with positive estimates and $p < 0.05$ using the DGRP correction when increasing the cut-off from 5% to 15%, even though it yields to lower α values. Also, using the DGRP correction and the cut-off set at 15%, we obtain a number of genes with $\alpha > 0$ and $p < 0.05$ similar to the FWW correction. Finally, the standard MKT leads to the lowest α estimates and number of genes with positive α and $p < 0.05$ among the 236 genes analyzed in this case.

Overall, the DGRP correction is the method that performed the best for detecting positive selection at the single-gene level. Contrary to other methods such as the FWW or the asymptotic MKT, on which some genes are not analyzable, this correction allows estimating α in the same number of genes than the Standard MKT. In addition, it detects the highest number of genes with positive α values and $p < 0.05$. Specifically, setting the cut-off at $\sim 5\%$ provides the best results. FWW correction also performs well, but the loss of power can be very dramatic if data has a large fraction of low frequency variants. Finally, the asymptotic approach is not meant to be used for single-gene analyses, as it requires a huge amount of polymorphic sites at different frequencies to perform. However, in the cases it can be applied, this method provides the highest unbiased α estimates.

Table 3.9: Genes analyzed by FWW method. Number of analyzable genes together with their summary adaptation metrics for each cut-off.

Cut-off	**Genes analyzed**	α **mean (± SD)**	**Genes** $\alpha > 0$ **and** $p < 0.05$	α **mean (± SD)**
0	11005	1.275 (± 3.956)	739	0.789 (± 0.138)
0.025	10249	-0.215 (± 2.559)	1039	0.861 (± 0.123)
0.075	9930	-0.043 (± 2.191)	996	0.876 (± 0.119)
0.125	9722	0.048 (± 1.855)	929	0.883 (± 0.113)
0.175	9545	0.100 (± 1.732)	881	0.887 (± 0.114)
0.225	9377	0.140 (± 1.701)	831	0.891 (± 0.112)
0.275	9214	0.179 (± 1.616)	774	0.901 (± 0.104)
0.325	9042	0.207 (± 1.576)	724	0.905 (± 0.105)

Table 3.10: Genes analyzed by DGRP method. Number of analyzable genes and their summary adaptation metrics for each cut-off.

Cut-off	Genes analyzed	α mean ($\pm$ SD)	Genes $\alpha > 0$ and $p < 0.05$	α mean ($\pm$ SD)
0	11005	1.275 ($\pm$ 3.956)	739	0.789 ($\pm$ 0.138)
0.025	11005	-0.735 ($\pm$ 3.615)	1102	0.819 ($\pm$ 0.130)
0.075	11005	-0.779 ($\pm$ 3.708)	1056	0.816 ($\pm$ 0.131)
0.125	11005	-0.809 ($\pm$ 3.693)	1011	0.814 ($\pm$ 0.130)
0.175	11005	-0.837 ($\pm$ 3.631)	989	0.810 ($\pm$ 0.132)
0.225	11005	-0.860 ($\pm$ 3.605)	974	0.808 ($\pm$ 0.132)
0.275	11005	-0.896 ($\pm$ 3.669)	948	0.804 ($\pm$ 0.134)
0.325	11005	-0.918 ($\pm$ 3.661)	928	0.803 ($\pm$ 0.134)

Table 3.11: Genes analyzed by Asymptotic MKT. Adaptation metrics of the 236 genes analyzable using the asymptotic MKT method computed with standard MKT, DGRP and FWW corrections.

	Asymptotic MKT	Standard MKT	DGRP 0.05	DGRP 0.15	FWW 0.05	FWW 0.15
Mean ($\pm$ SD)	0.774 ($\pm$ 0.198)	0.714 ($\pm$ 0.139)	0.773 ($\pm$ 0.139)	0.763 ($\pm$ 0.150	0.817 ($\pm$ 0.167)	0.840 ($\pm$ 0.165)
$\alpha > 0$ and $p < 0.05$	35	14	18	83	84	81

ii. Genome-scale estimates of adaptive evolution

Here we compare α estimates obtained using each of the MKT-derived methods on the complete set of 13,753 protein-coding genes for the six *D. melanogaster* populations.

In detail, polymorphism and divergence metrics of all genes for each population were grouped together, creating a sort of "concatenated" gene on which the diverse tests were applied. In these analyses the number of polymorphic sites is large enough for fitting the exponential model in the asymptotic MKT. This is true in all cases except the chromosome arm 3R of the Asian meta-population, on which the linear fitting is used.

Table 3.12 shows the α estimates for each meta-population considering genes from: (i) the complete genome, $n = 13,753$; (ii) autosomes, $n = 11,584$; (iii) chromosome arms 2L, $n = 2,647$; (iv) 2R, $n = 2,822$; (v) 3L, $n = 2,723$; (vi) 3R, $n = 3,392$ and; (vi) X chromosome, $n = 2,169$. Only polymorphic sites with frequencies in the range [0, 0.9] were used in the asymptotic method, and we considered a cut-off of 10% in the DGRP and FWW corrections because polymorphic sites of Asia and Oceania meta-populations were classified only in 10 DAF categories due to the low number of samples that they include (18 and 15 individuals, respectively). We are aware that this limitation of both the data and the method may cause a slightly underestimation of α in the DGRP correction test, as we previously demonstrated, but we do not expect any significant bias which could cause a misinterpretation of the major results.

First, we observe that asymptotic MKT provides the highest estimates of α, followed by FWW and DGRP corrections, while standard MKT underestimates α in all situations.

Second, results show that differences in adaptation rates between chromosomes are similar in all meta-populations. Specifically $\alpha_{\ X} > \alpha_{\ 2R} > \alpha_{\ 2L} > \alpha_{\ 3L} > \alpha_{\ 3R}$. As shown in Table 3.12, the X/A ratio is higher than one in all meta-populations and tests. Differences between X chromosome and autosomes adaptation rates are very pronounced using the standard MKT (ratios ranging from ~ 1.7 to ~ 2.6) but lower for the methods that account for the presence of slightly deleterious mutations. Indeed, X/A rates obtained with the asymptotic MKT are the lowest, but still > 1 (ratios ~ 1.3).

Third, we compared visually the rate of adaptive evolution among the six *D. melanogaster* meta-populations using results obtained with the asymptotic MKT. We observe that Asia shows the lowest values. Then, Europe / N. Africa, America and Oceania show very similar rates, with the order of α values varying in each chromosome arm. In addition, there is also a latitudinal African cline, with Souther Africa showing higher adaptation rates than Equatorial Africa meta-populations. This is the same trend observed previously for nucleotide diversity rates,

In summary, results show that, on average, more than 60% of fixed differences are adaptive in the genome of *D. melanogaster* relative to *D. simulans*, with α values estimated with the asymptotic MKT ranging from ~ 0.60 for Asia samples, ~ 0.64 for Oceania, America and Europe / N. Africa meta-populations and up to ~ 0.68 for African individuals (Figure 3.19). In addition, X chromosome presents higher α values in all cases, ranging from ~ 0.75 to ~ 0.83.

Table 3.12: Adaptive evolution in the six *Drosophila* meta-populations. Adaptation metrics (α) for the asymptotic test were estimated using polymorphism values with frequencies between 0 and 0.9. *G: Genome, A: Autosomes; *: linear fitting in the asymptotic MKT.*

		Asia	**Oceania**	**America**	**Europe / N. Africa**	**Equat. Africa**	**Southern Africa**
G	Standard	0.44769	0.47783	0.41322	0.40447	0.36683	0.48069
n=13,753	FWW 10%	0.49851	0.55363	0.57125	0.57811	0.62687	0.64212
	DGRP 10%	0.49153	0.53664	0.52104	0.51876	0.4910	0.5641
	Asymptotic	0.60471	0.64951	0.63514	0.63502	0.68148	0.69083
A	Standard	0.37692	0.40602	0.32891	0.32439	0.25753	0.39272
n=11,584	FWW 10%	0.43033	0.49078	0.51094	0.52108	0.56791	0.58181
	DGRP 10%	0.42351	0.47167	0.45261	0.45313	0.40934	0.49206
	Asymptotic	0.55284	0.60270	0.58562	0.58891	0.63247	0.64255
2L	Standard	0.42265	0.46153	0.4118	0.4089	0.34875	0.47703
n=2,647	FWW 10%	0.51498	0.55212	0.56557	0.57152	0.61123	0.62982
	DGRP 10%	0.49418	0.53188	0.51902	0.52016	0.48296	0.55606
	Asymptotic	0.57223	0.63736	0.62616	0.63961	0.66832	0.67081
2R	Standard	0.40423	0.44519	0.37524	0.37121	0.30826	0.42171
n=2,822	FWW 10%	0.51084	0.53463	0.54829	0.55526	0.59608	0.60348
	DGRP 10%	0.48711	0.51346	0.49107	0.49399	0.44656	0.51916
	Asymptotic	0.59357	0.63180	0.62224	0.61499	0.65645	0.66627
3L	Standard	0.37748	0.38951	0.30429	0.29946	0.23705	0.36016
n=2,723	FWW 10%	0.37785	0.48049	0.49469	0.50977	0.56515	0.57709
	DGRP 10%	0.37785	0.45989	0.43351	0.4342	0.3957	0.47457
	Asymptotic	0.47550	0.59838	0.58052	0.58311	0.62596	0.63144
3R	Standard	0.29493	0.3237	0.21572	0.21515	0.13294	0.30783
n=3,392	FWW 10%	0.2945	0.3878	0.42928	0.44571	0.49926	0.51635
	DGRP 10%	0.29456	0.37396	0.35919	0.35922	0.30875	0.41666
	Asymptotic	0.40496*	0.53683	0.50807	0.51198	0.57952	0.60707
X	Standard	0.64498	0.68695	0.66487	0.6335	0.67199	0.72513
n=2,169	FWW 10%	0.69652	0.73905	0.74771	0.7412	0.79458	0.81311
	DGRP 10%	0.68543	0.72797	0.72341	0.70841	0.72292	0.76658
	Asymptotic	0.75088	0.78094	0.77683	0.76322	0.81907	0.82981
X/A	Standard	1.71118	1.69191	2.02143	1.95289	2.60936	1.84643
	FWW 10%	1.61857	1.50587	1.4634	1.42243	1.39913	1.39755
	DGRP 10%	1.61845	1.54339	1.59831	1.56337	1.76606	1.5579
	Asymptotic	1.35818	1.29567	1.3265	1.29597	1.29507	1.29152

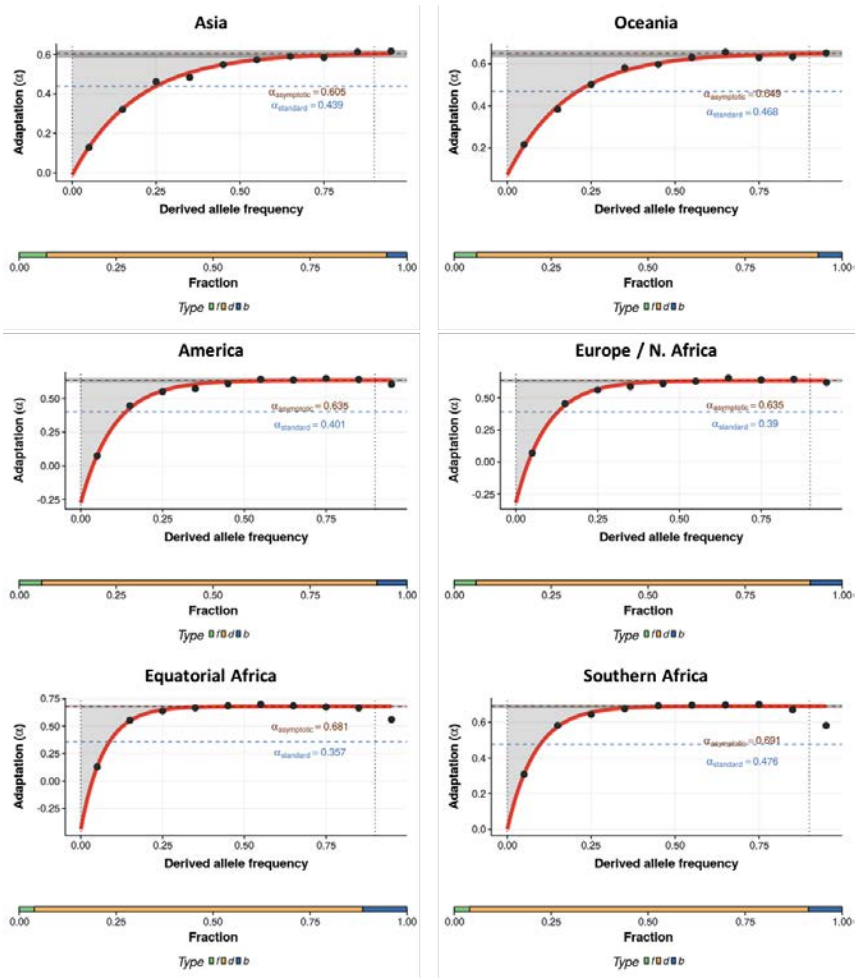


Figure 3.19: Adaptive and purifying selection in six *Drosophila* meta-populations using integrative MKT. Graphical representation of (i) the rate of adaptive evolution (α) and the fraction of sites under purifying selection (d: strongly deleterious, b: weakly deleterious, f: neutral) estimated using the asymptotic approach iMKT for the six *D. melanogaster* meta-populations. Note that values in the Y-axis vary between graphs. Adaptation metrics are higher in African than in non-African samples, and Asia meta-populations shows the lowest α estimate. The shaded gray area in the top graphs corresponds to the fraction of weakly deleterious (b) sites, which lower the estimate of α at low frequencies in all situations.

iii. Quantifying the fraction of purifying selection along the genome

We estimated the fraction of putatively selected sites which are strongly deleterious (d), weakly deleterious (b) and neutral (f) using both the DGRP correction and the integrative MKT for the six *D. melanogaster* meta-populations. As in the previous section, we used a cut-off of 10% for the DGRP method due to the limitation in the number of samples of certain populations, and again, the chromosome 3R of the Asian population was not analyzable using the iMKT (data can not fit an exponential model). We considered again (i) the complete set of genes; (ii) only autosomic genes and; (iii) each chromosome arm independently (Table 3.13).

The fraction of strongly deleterious sites (d) is very similar in both methodologies, as it is based in the same theoretical assumption, but we observe that integrative MKT is able to detect a larger fraction of weakly deleterious mutations (b) than DGRP. Contrary, the fraction of neutral sites (f) is larger using the DGRP correction than the iMKT, maybe because some deleterious variants are wrongly assigned to be neutral based on their frequency and the arbitrary cut-off applied in the DGRP test. Probably, a better adjustment of the frequency cut-off applied in the DGRP method would lead to estimates closer to the ones obtained with the iMKT.

Here, we focus the description of results obtained using the iMKT in order to be consequent with the previous and following sections. Figure 3.19 shows the graphical representation of the integrative MKT method, with (i) the estimation of α and (ii) the fraction of sites under purifying selection.

In general, $\sim 84 - 87\%$ of sites appear to be strongly deleterious (d), $\sim 4 - 7\%$ neutral (f) and $\sim 5 - 11\%$ weakly deleterious (b) (Figure 3.19). There are not clear differences between chromosomes, but chromosome arm 3R shows a larger fraction of slightly deleterious mutations than the rest in all meta-populations. Regarding the X/A ratio, in general we do not observe clear differences between sexual and autosomic chromosomes. However, the fraction of b is slightly lower in the sexual chromosome in all meta-populations (X/A ratios $\sim 0.82 - 0.88$) except Asia (1.17345).

Table 3.13: Purifying selection fractions in the six *Drosophila* meta-populations. Again we only considered polymorphism values with frequencies between 0 and 0.9. *G: Genome, A: Autosomes; d: strongly deleterious; b: weakly deleterious; f: neutral.*

			Asia	Oceania	America	Europe / N. Africa	Equatorial Africa	Southern Africa
G n=13,753	DGRP 10%	d	0.87698	0.88061	0.86416	0.86144	0.84661	0.87330
		f	0.11325	0.10594	0.11088	0.11196	0.12331	0.10635
		b	0.00976	0.01345	0.02496	0.02659	0.03008	0.02035
	iMKT	d	0.87698	0.88061	0.86416	0.86144	0.84661	0.87330
		f	0.07061	0.05800	0.05846	0.05661	0.03931	0.03974
		b	0.05240	0.06139	0.07738	0.08194	0.11408	0.08696
A n=11,584	DGRP 10%	d	0.87765	0.88023	0.86307	0.86137	0.84275	0.87050
		f	0.11320	0.10653	0.11169	0.11221	0.12510	0.10831
		b	0.00915	0.01324	0.02524	0.02642	0.03215	0.02118
	iMKT	d	0.87766	0.88023	0.86307	0.86137	0.84275	0.87050
		f	0.07104	0.05755	0.05828	0.05592	0.04063	0.04073
		b	0.05131	0.06222	0.07865	0.08271	0.11662	0.08877
2L n=2,647	DGRP 10%	d	0.87793	0.88270	0.86952	0.86739	0.84921	0.87750
		f	0.10695	0.10197	0.10670	0.10764	0.11972	0.10399
		b	0.01512	0.01533	0.02379	0.02496	0.03107	0.01851
	iMKT	d	0.87793	0.88270	0.86952	0.8674	0.84921	0.87750
		f	0.06662	0.05945	0.05883	0.0568	0.04198	0.04048
		b	0.05545	0.05784	0.07165	0.0758	0.10881	0.08202
2R n=2,822	DGRP 10%	d	0.87828	0.88439	0.86842	0.86784	0.84872	0.87347
		f	0.10479	0.10138	0.10719	0.10636	0.12103	0.10521
		b	0.01693	0.01423	0.02439	0.02580	0.03024	0.02132
	iMKT	d	0.87827	0.88439	0.86842	0.86784	0.84872	0.87347
		f	0.06429	0.05603	0.05451	0.05490	0.03850	0.04034
		b	0.05744	0.05958	0.07707	0.07726	0.11278	0.08619
3L n=2,723	DGRP 10%	d	0.88349	0.88241	0.86450	0.86260	0.84509	0.86950
		f	0.11644	0.10403	0.11033	0.11097	0.12270	0.10717
		b	0.00007	0.01356	0.02517	0.02643	0.03222	0.02334
	iMKT	d	0.88349	0.88241	0.86450	0.86260	0.84509	0.86950
		f	0.11023	0.05635	0.05654	0.05325	0.03914	0.04069
		b	0.00628	0.06124	0.07895	0.08415	0.11577	0.08981
3R n=3,392	DGRP 10%	d	0.86892	0.87081	0.84884	0.84786	0.82780	0.86121
		f	0.13116	0.11959	0.12351	0.12421	0.13729	0.11697
		b	-0.00008	0.00961	0.02765	0.02793	0.03491	0.02182
	iMKT	d	-	0.87081	0.84884	0.84786	0.82780	0.86121
		f	-	0.05679	0.06357	0.05840	0.04275	0.04105
		b	-	0.07240	0.08759	0.09374	0.12945	0.09775
X n=2,169	DGRP 10%	d	0.87269	0.88532	0.87598	0.86468	0.86758	0.88822
		f	0.11282	0.09964	0.10235	0.10767	0.11186	0.09492
		b	0.01450	0.01504	0.02166	0.02765	0.02056	0.01685
	iMKT	d	0.87268	0.88533	0.87598	0.86467	0.86758	0.88822
		f	0.06712	0.06309	0.05938	0.06252	0.03299	0.03462
		b	0.06021	0.05157	0.06464	0.07281	0.09943	0.07716
X/A	DGRP 10%	d	0.99435	1.00578	1.014958	1.00384	1.02946	1.02036
		f	0.99664	0.93532	0.91637	0.95954	0.89416	0.87637
		b	1.58469	1.13595	0.85816	1.04655	0.63950	0.79556
	iMKT	d	0.99432	1.00579	1.01496	1.00383	1.02946	1.02036
		f	0.94482	1.09626	1.01887	1.11803	0.81196	0.84999
		b	1.17345	0.82883	0.82187	0.88030	0.85259	0.86921

Then, when we compare visually the values obtained for each meta-population we observe that the fraction of strongly deleterious sites (d) from lowest to highest estimates corresponds to: Equatorial Africa, Europe / N. Africa, America, Souther Africa, Asia and Oceania. Then, the fraction of weakly deleterious sites (b) is lower in Asia, followed by Oceania, America, Europe / N. Africa, Southern Africa and finally the highest estimate corresponds to Equatorial Africa samples. Finally, the fraction of neutral sites (f) follows the opposite pattern than b. This tendency is maintained in all autosomic arms. However, for the X chromosome Oceania shows a lower fraction of weakly deleterious variants than Asia ($b = 0.05157$ and $b = 0.06021$, respectively).

3.3.4 The effect of recombination and coding density in the rates of nucleotide variation and adaptation

One of the major goals of evolutionary biology is the identification and description of the genomic determinants of nucleotide variation and, therefore, of molecular evolution. Taking advantage of the huge amount of genomic information obtained in this work, we analyzed the relationship between some genomic determinants of genetic change such as **recombination** and **coding density** and (i) the observed patterns of nucleotide variation at intra- and inter-species level (polymorphism and divergence, respectively) and (ii) the rate of adaptive and purifying selection in the *D. melanogaster* genome.

We did not analyze the influence of mutation rate (μ), another well-known determinant of the rate of genetic variation, because we do not have the adequate data to do so. The best *proxy* for the mutation rate that we can use is the rate of synonymous divergence (k_0), but these estimates are not independent from the levels of divergence (k) and adaptation (α) and therefore, analyzing the correlation between such variables may lead to highly biased results.

First, we used 100 kb windows-based estimates of population historical recombination (ρ), coding density (i.e., the proportion of coding nucleotides relative to the total number of sites in the window), nucleotide diversity (π) and divergence with *D. simulans* (k) to analyze the correlations among these parameters.

Figure 3.20 shows the Spearman's correlation coefficients (Spearman 1904) between each pairwise comparison of metrics for each meta-population considering whole genome data. We observe that historical recombination is positively correlated with polymorphism (correlation coefficients between 0.52 for the Oceania population and 0.77 for the Equatorial Africa meta-population; $p < 0.05$). and negatively correlated with divergence (correlation coefficients in the range from -0.15 to -0.22; $p < 0.05$). Besides, the correlation coefficients between recombination (ρ) and divergence (k) pinpoint a negative association in all populations (coefficients from -0.24 to -0.11; $p < 0.05$) except Southern Africa, where we observe a positive estimate of 0.06, but with an associated $p = 0.0486$.

Besides, there is a weak and negative correlation between coding density and π (correlation coefficients of -0.07 and -0.09 for America and Southern Africa meta-populations, respectively; $p < 0.05$) and a weak and positive correlation between coding density and divergence (coefficients between 0.08 and 0.09; $p < 0.05$) for some meta-populations. Recombination and coding density also show a positive and weak correlation with coefficients of 0.06 and 0.07 for Asia and Oceania populations, respectively; $p < 0.05$. Finally, there is a negative correlation between polymorphism and divergence rates (coefficients around -0.4 for non-African populations and somewhat lower for African populations, with values of -0.27 for Equatorial Africa and -0.19 for Southern Africa; $p < 0.05$).

Results regarding each chromosome arm independently can be accessed at Supplementary Figures 7.2 and 7.3. We observe in every chromosome the same patterns described for full-genome metrics. In addition, the distribution of ρ and π values shown in Supplementary figure 7.2 suggests that an asymptotic model would fit the data better than a linear one, which highlights the presence of an optimal recombination threshold, as suggested by Mackay et al. (2012), Castellano et al. (2016) and Barrón (2015). Correlation coefficients between ρ and polymorphism in the X chromosome are in general lower than in the autosomes, but the asymptotic correlation is maintained. We also observe lower correlation coefficients for the X chromosome than the autosomes between ρ and divergence metrics.

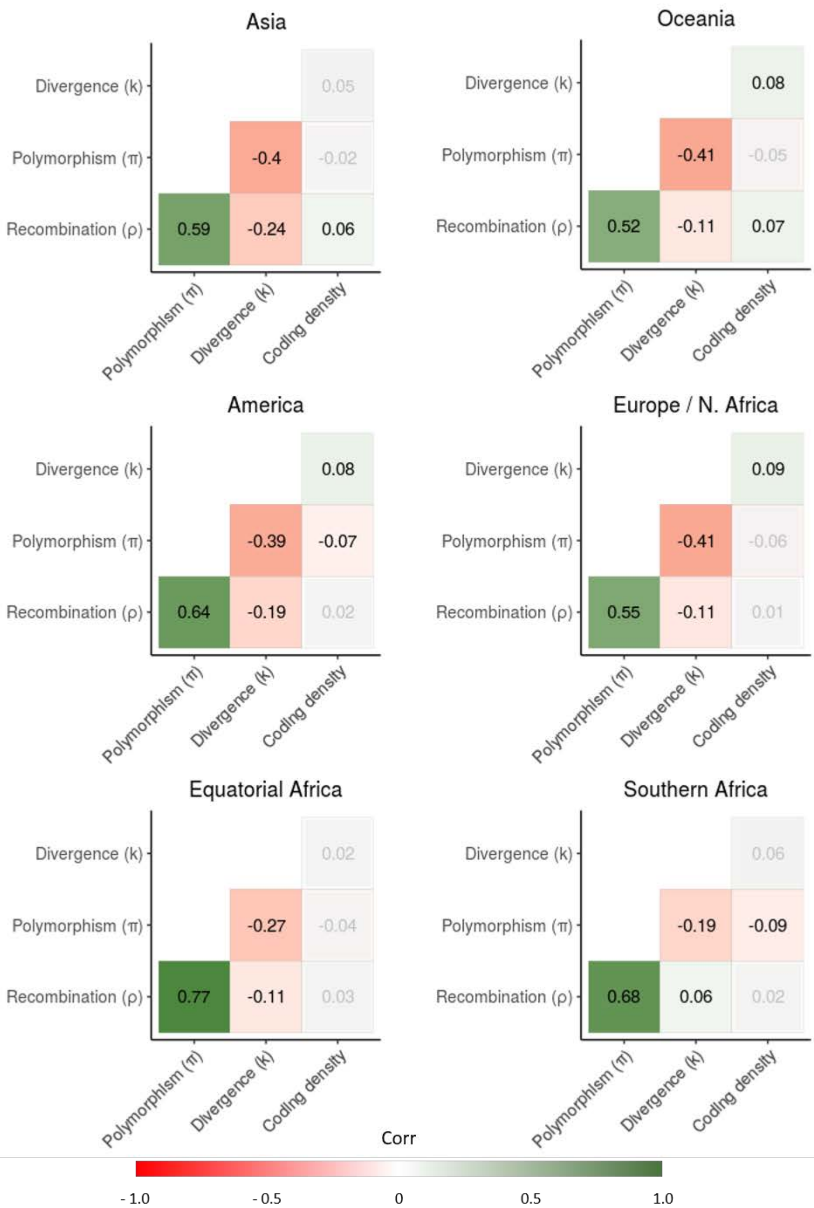


Figure 3.20: Spearman's correlation coefficients between nucleotide variation metrics and genome variables. Historical recombination (ρ), coding density, nucleotide diversity (π) and divergence with *D. simulans* (k) metrics estimated in 100 kb windows for each meta-population. Correlation coefficients with an associated $p > 0.05$ are shaded in grey.

Second, we used gene-based estimates to analyze pairwise correlations between ρ, coding density, the rate of adaptive evolution (α) and the fractions of sites under purifying selection (d: strongly deleterious, b: weakly deleterious, f: neutral).

Specifically, we grouped genes in ten equally sized bins (of 1,375 genes each) according to (i) recombination and (ii) coding density values and we applied the integrative MKT on each of these bins to obtain adaptation metrics. Adaptation and recombination estimates for ten bins of ρ and each meta-population can be accessed at Supplementary table 7.1. The list of bins based on coding density can be accessed at Supplementary Table 7.2.

Figure 3.21 shows Spearman's correlation coefficients estimated using the ten bins according to population recombination (ρ) for each meta-population considering the complete set of protein-coding genes. Results corresponding to coding density bins can be accessed at Supplementary figure 7.4.

Results show that historical recombination is positively correlated with both the rate of adaptive evolution (correlation coefficients between 0.77 for Oceania population and 1 for Europe / North Africa meta-population; $p < 0.05$) and the fraction of strongly deleterious sites (d, coefficients in the range from 0.88 to 0.99; $p < 0.05$). There is also a negative correlation between ρ and the fractions of neutral (f) and weakly deleterious (b) sites (correlation coefficients from -0.76 to -0.92 and from -0.9 to -1, respectively; $p < 0.05$). We also see that adaptation (α) is positively correlated with d (coefficients from 0.71 to 0.99; $p < 0.05$) and negatively correlated to f and b (coefficients from -0.77 to -0.98 and from -0.89 to -0.99, respectively; $p < 0.05$). Finally, we observe a strong and negative correlation between d and f (coefficients in the range from -0.65 to -0.95; $p < 0.05$) and between d and b (coefficients from -0.98 to -1; $p < 0.05$) and a positive correlation between f and b (coefficients from 0.83 to 0.94; $p < 0.05$).

Then, regarding coding density bins, we observe that there are no correlation coefficients with $p < 0.05$ and thus we can state that both parameters are not correlated. Indeed, coding density appears to be only negatively correlated to the fraction of neutral sites (coefficients from -0.65 to -0.78; $p < 0.05$). In the other comparisons using the proportion of coding sites as one of the parameters we do not

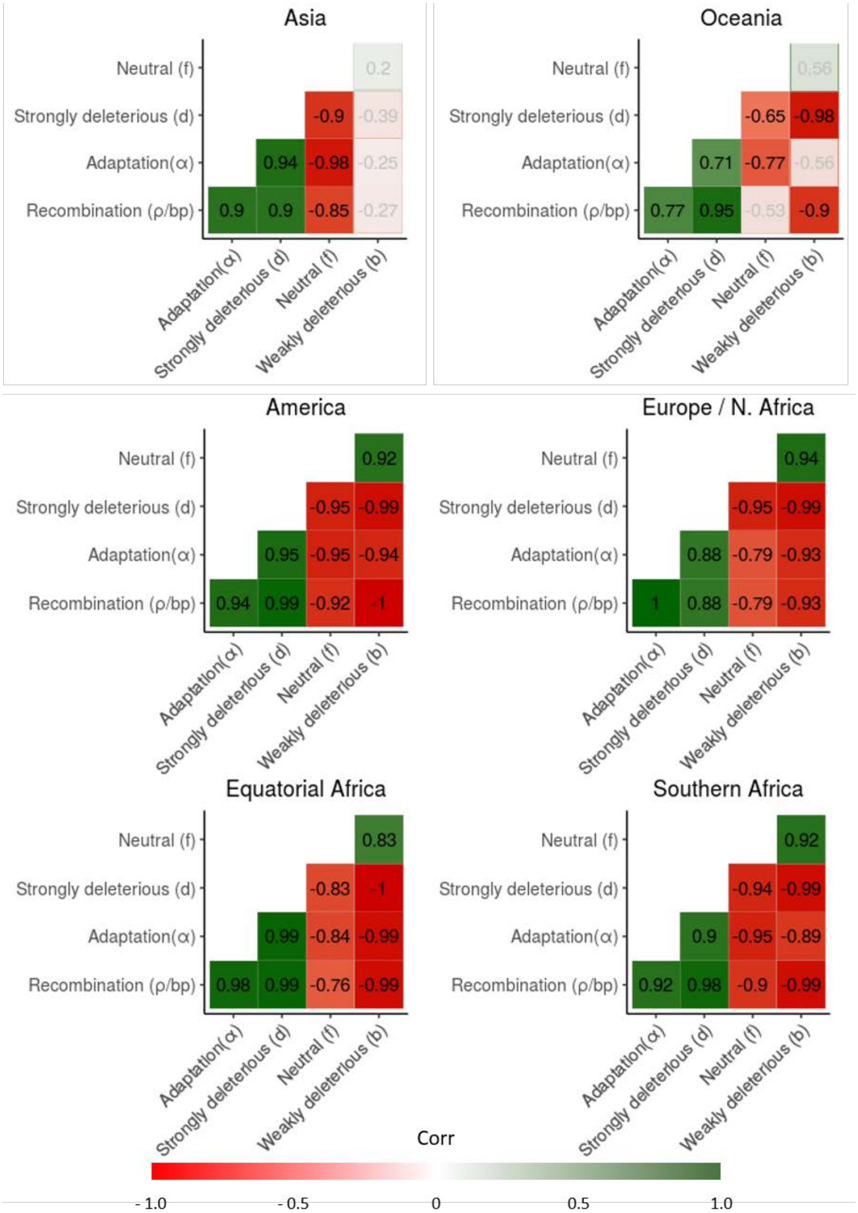


Figure 3.21: Spearman's correlation coefficients between adaptation metrics and recombination. Recombination (ρ), rate of adaptive evolution (α) and fraction of strongly deleterious (d), neutral (f) and weakly deleterious sites (b) for 13,754 genes grouped in ten bins based on ρ estimates for each meta-population. Correlation coefficients with an associated $p > 0.05$ are shaded in grey.

get statistical support (i.e., correlation coefficients with $p < 0.05$). Adaptation metrics (α) only appear to be negatively correlated with f in the Asia population (correlation coefficient of -0.77 with a $p = 0.03$). Then, we observe the same trend as before regarding the correlations between the diverse fractions of sites under purifying selection. In detail, we see a negative correlation between d and both f and b, with correlation coefficients ranging from -0.77 to -0.93 and from -0.64 to -0.96, respectively; $p < 0.05$). Finally, the correlation between f and b is positive as described above (coefficients of 0.76 for America and Equatorial Africa and 0.87 for Southern Africa meta-populations; $p < 0.05$).

Figure 3.22 shows the α and b estimates for each bin of recombination and coding density. We clearly observe a positive non-linear correlation between ρ and α and a negative non-linear correlation between ρ and b. Besides, the relationship between coding density and both metrics does not appear to follow any clear pattern.

Interestingly, ρ and α metrics follow an asymptotic distribution (similar to the one observed between ρ and π) which again highlights the presence of an optimal threshold of recombination (r_{opt}) above which sites segregate independently and therefore the adaptation rate is not affected by linked selection (Castellano et al. 2016, Casillas & Barbadilla 2017). Thus, we used an adaptation of the approach applied to estimate the fraction of b in the integrative MKT, in order to estimate the reduction in adaptation rate caused by the Hill-Robertson interference (L_{HRi}) and the optimal recombination value (r_{opt}) in the six *D. melanogaster* meta-populations (Table 3.14).

We observe that there are not clear differences between meta-populations. In detail, Equatorial Africa and America meta-populations show the largest rate of L_{HRi} ($\sim 24\%$) followed by Asia, Southern Africa, Europe/N. Africa and Oceania, with rates ranging from $\sim 19.5\%$ to $\sim 21.3\%$. Castellano et al. (2016) found that L_{HRi} on the Raleigh North American population (RAL) was on average 27%, an estimate which is somehow close to the value obtained in this work (24%). On average, results show that $L_{HRi} \sim 22\%$ in the genome of *D. melanogaster*, considering all six meta-populations together.

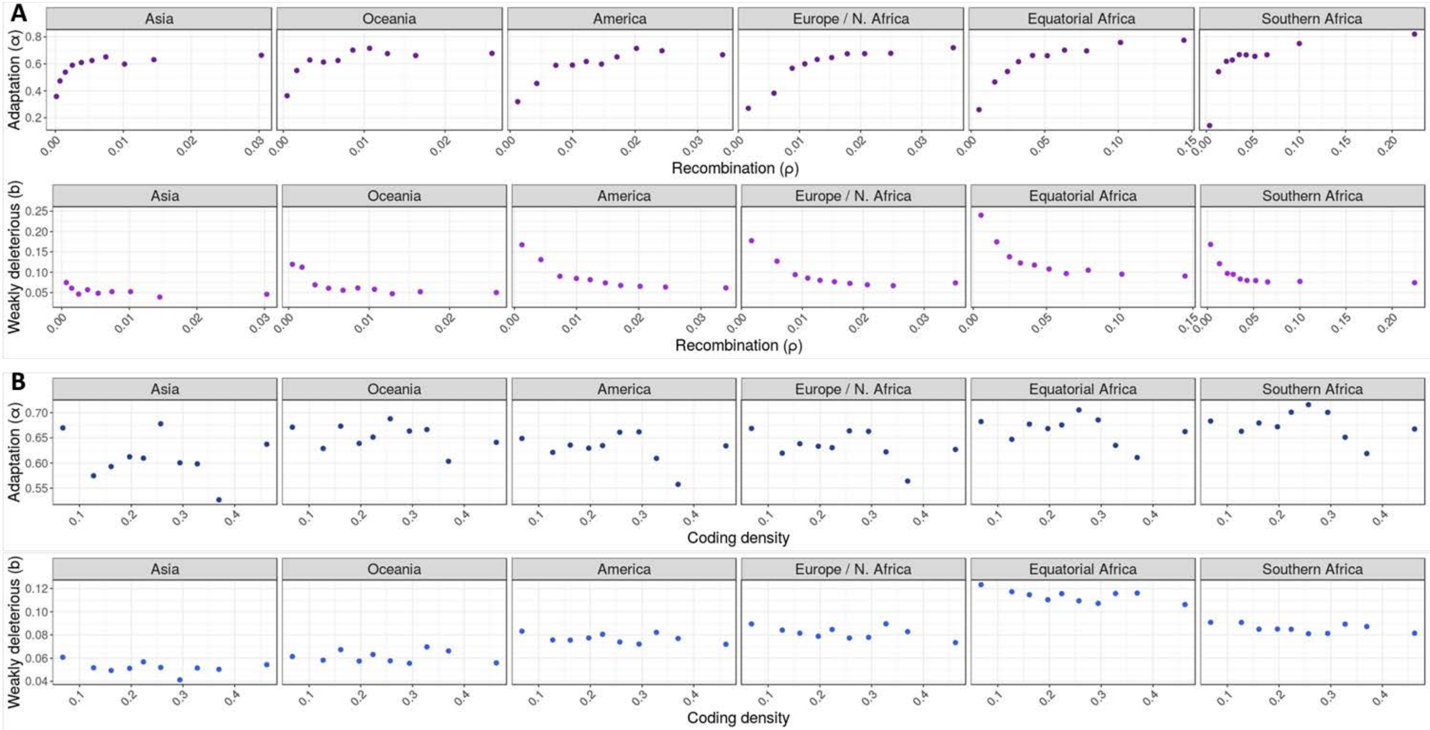


Figure 3.22: Graphical representation of α and b in bins of genes grouped by ρ and coding density. (A) We observe a positive association between recombination and the rate of adaptation (top), while it is negatively associated with the fraction of weakly deleterious sites (bottom). **(B)** Coding density does not seem to be affecting neither the rate of adaptive evolution nor the proportion of sites with slightly deleterious fitness effects.

Table 3.14: Hill-Robertson interference in six *D. melanogaster* populations. Results correspond to the complete set of 13,754 protein coding genes grouped in 10 bins based on ρ metrics.

	Asia	**Oceania**	**America**	**Europe / N. Africa**	**Equatorial Africa**	**Southern Africa**
Recombination optimal (r_{opt})	0.00539	0.00676	0.01463	0.01526	0.07852	0.06529
Hill-Robertson Load (L_{HRi})	0.21339	0.19555	0.24085	0.20969	0.24226	0.21319

4. Discussion

4. Discussion

4.1 Bioinformatics tools for population genomics

The increasing amount of genomic data available nowadays, caused by the lowering of the cost and improvement of new sequencing technologies, requires the continuous development and optimization of bioinformatics tools able to handle and analyze such information. In this regard, the implementation of new biologically-oriented software is constantly growing (Rigden & Fernández 2017), a tendency that is not predicted to decelerate in the next years (Li et al. 2016).

Bioinformatics tools might help addressing several steps from data acquisition, filtering, processing, visualization or analysis to the final reporting step. This work focuses on the development and application of two tools for dealing with two of the steps stated previously: (i) the visualization and (ii) the analysis of genetic data, both from a population genomics perspective.

Results presented in this work focus on analyses performed using the *Drosophila Genome Nexus* project data (DGN, Lack et al. 2015, 2016), but furthermore, the two tools developed during this project were also applied to human data from the 1000 Genomes Project (1000GP, Consortium et al. 2015), even that these analyses are out of the scope of this thesis. Briefly, we first developed PopFly (Hervas et al. 2017), a population genomics oriented web browser with information from the DGN project, which lead to the development of PopHuman, the human population genomics browser (Casillas et al. 2018). Second, we implemented a battery of MKT-derived methods in an R package named iMKT, which lead to the development of a web-server that includes all the functionalities from the iMKT R package. Thus, in the following sections, we first discuss the strengths and weaknesses of the PopFly novel genome browser and then we shift to to review the integrative MKT software.

4.1.1 PopFly: the *Drosophila* population genomics browser

With the launch of large-scale sequencing projects such as the DGN (Lack et al. 2015, 2016) and the 1000GP (Consortium et al. 2015), the need of tools able to properly store and allow a graphical representation of such information has become a pivotal step in population genetics studies. Indeed, a visual display of the estimated metrics describing genome-wide variation and selection patterns would allow gaining a global view and understanding of the evolutionary forces shaping genome variation.

Genome browsers are precisely designed to face this challenge, and their development and improvement is continuously growing as the amount of available sequencing data does too. In the last years, next generation genome browsers have implemented some novel features in order to deal with such large data sets, while enhancing the user's final experience when using the tool. In this regard, the JBrowse software (Buels et al. 2016) has emerged as a very valuable tool for building up custom genome browsers, as it allows a very easy implementation of the whole framework in a web server, making all information available and easy to share. Indeed, most species-specific genome browsers (FlyBase, WormBase or ZFIN) that were initially based on the GBrowse framework (Stein et al. 2002) are now shifting to JBrowse (*see 1.4.1 Genome browsers: graphical biological databases*), which demonstrates its suitability for dealing with this type of data.

In addition, web services are becoming a standard protocol for data exchange and application communication in all knowledge areas, including biology. However, the problem of how to define the data exchange format and the application interface is still unsolved. Most genome browsers (and bioinformatics application platforms in general) are gradually moving to cluster servers or cloud environments, which appears to be an adequate solution to avoid heavy data transmission (Wang et al. 2013). For instance, Ensembl (Flicek et al. 2009) and JBrowse are actively using Amazon web services to improve the on-line service; and in the future, we expect that more and more cloud technologies will provide high performance for the end users.

Overall, with the development of new sequencing and web technologies, together with the increasing number of population genomics projects performed in the last years, genome browsers have become a key collaboration platform for researchers to share data and to exchange knowledge (Nielsen et al. 2010). From a population genomics perspective, the development of specific populations-oriented genome browsers able to allow an easy retrieval and visualization of such data is one the major bioinformatics challenges right now.

In this regard, we developed PopFly, a population genomics-oriented genome browser, based on JBrowse software, that contains a complete inventory of population genomic parameters estimated from the *Drosophila Genome Nexus* data (Lack et al. 2015, 2016).

This browser is designed for the automatic analysis and display of genetic variation data within and between populations along the *D. melanogaster* genome, allowing the visualization and retrieval of functional annotations, estimates of nucleotide diversity metrics, linkage disequilibrium statistics, recombination rates, a battery of neutrality tests, and population differentiation parameters at different window sizes through the euchromatic chromosomes. PopFly is open and freely available at site `https://popfly.uab.cat`; and it has been designed to work in the most used web browsers (Chrome, Firefox, Safari, Microsoft Edge, Internet Explorer and Opera).

This new genome browser is based on a similar instance previously developed by our group named PopDrowser (Ràmia et al. 2011). The latter was built up as part of the Drosophila Genetic Reference Panel project (DGRP, Mackay et al. 2012) and it hosts a catalog of population genomics statistics for one single *D. melanogaster* population from Raleigh, North America.

Compared to PopDrowser, PopFly presents two significant advantages. First, data is not limited to a single population. This allows detecting very recent selective sweeps that have occurred in a single population, or older selective sweeps shared among a few related populations, whose detection gives a reinforcement of the time depth and biology underlying the specific selection signal (Casillas et al. 2018). Second, PopFly is based on JBrowse, whereas PopDrowser is implemented in a GBrowse framework, the previous genome browser software version developed by the GMOD community (Stein et al. 2002). While in most existing genome browsers architectures, such as GBrowse, the genome is rendered into images on the web-server and the role of the client is restricted to displaying those images, JBrowse distributes work between the server and client and therefore uses significantly less server overhead than previous genome browsers. Moreover, JBrowse helps to preserve the user's sense of location by avoiding discontinuous transitions, offering instead smoothly animated panning, zooming, navigation, and track selection (Skinner et al. 2009). This allows PopFly to outperform its previous version in terms of speed and execution while providing a better user experience when navigating the site.

In addition, the JBrowse open-source framework nature allowed us to (i) modify parts of the code to upgrade some built-in functions and adapt their functionality to the huge amount of data stored in this server, and (ii) implement new utilities and support resources to facilitate performing population genetics analyses and retrieving data (*see 3.1.2. Utilities and support resources*).

Since its public release on June 2017, the PopFly genome browser has established as a reference tool for population genetics studies in *D. melanogaster*, as supported by the increasing traffic of users it has experienced since its release (*see Box 4.1*). Two key factors have driven this process: (i) the direct link established from FlyBase to PopFly (Figure 3.1); and (ii) the announcement of PopFly as the official repository of the Drosophila Genome Nexus Project through the DGN principal diffusion channels: DGN main website (`http://www.johnpool.net/genomes.html`) and the *Drosophila population genomics* Google group (`https://groups.google.com/forum/#!forum/drosophila-population-genomics`) with 230 members, created in 2008 and administered since then by Dr. John Pool (head of the DGN research project).

Currently (August 2018), PopFly has been cited in five scientific publications. In brief, researchers used PopFly to download sequence data, analyze nucleotide variation levels or examine adaptation estimates for genes and regions of interest. Three of these articles have been already published in high-impact journals (Casillas et al. 2018, Telonis-Scott & Hoffmann 2018, De Castro et al. 2018). The remaining two (Rech et al. 2018, Da Lage et al. 2018) are still published in bioRxiv (`https://www.biorxiv.org`), a free online archive and distribution service for unpublished preprints in the life

Box 4.1: PopFly users traffic

PopFly incorporates the Google Analytics functionality that allows tracking and collecting information of users, sessions, and so on. Since its public release on June 19, 2016 to August 19, 2018, the summary statistics are:

- A total of 2,758 users (106 users/month), accounting for 4,516 sessions (173 sessions/month), 6,619 pages (276 pages/month), and an average time spent of 01:47 minutes per session.

- The most used web browsers to access PopFly are: Chrome (1,604 users, 58.16%); Firefox (675 users, 24.47%); Safari (316 users, 11.46%); Microsoft Edge (81 users, 2.94%); Internet Explorer (49 users, 1.78%); and Opera (11 users, 0.4%).

- As shown in Figure 4.1A, the traffic of users is increasing and stabilizing over time. Note that we observe two peaks of traffic: May/June 2017 (when the PopFly research paper was initially published on-line) and January 2018 (coinciding with the MSc in Bioinformatics at UAB).

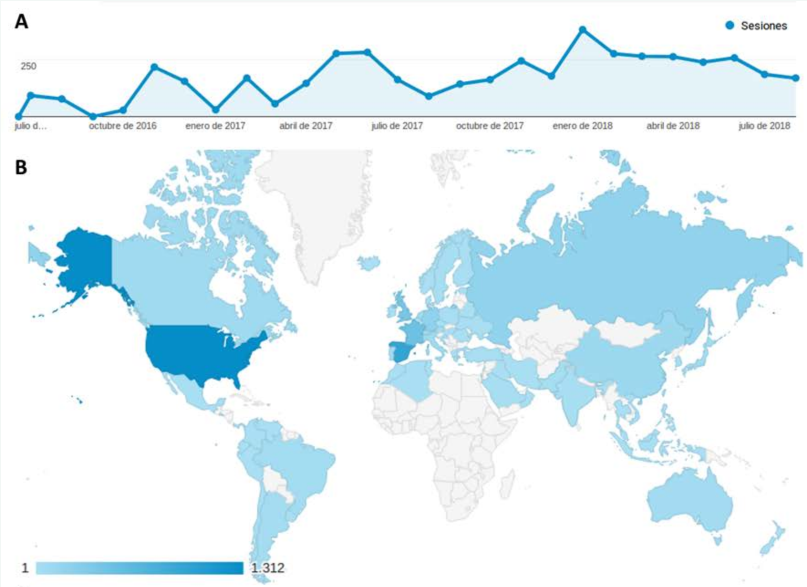
- If we consider only the last 6 months, statistics are: 1,005 users (167 users/month), 1,469 sessions (244 sessions/month) and 1,985 visited pages (331 pages/month), with an average time of 01:24 minutes per session. This shows that the increase of traffic is constant and not produced by the specific events mentioned above.

- Interestingly, geographic statistics (Figure 4.1B, C) show that PopFly is being accessed from all over the world, with most of its users residing in the United States of America (31.41% of users), followed by Spain (17.77%), United Kingdom (9.32%), France (8.63%), and Germany (5.04%).

- Finally, considering the 1,433 users and 1,985 sessions whose origin was tracked, we observe that users access PopFly from:

- FlyBase: 906 users (63.22%), 1,151 sessions (57.98%), with an average time of 00:37 minutes per session. FlyBase Beta: 28 users (1.95%), 43 sessions (2.17%), with an average of 01:52 minutes per session.
- DGN diffusion channels: 162 users (11.30%), 310 sessions (15.62%), with an average time of 03:35 minutes per session.

Box 4.1: (*Cont.*) PopFly users traffic



C

	Country	Acquisition			Behavior	
		Users	New Users	Sessions	Bounce Rate	Pages / Session
		0 % of Total: 0.00% (0)	2,475 % of Total: 100.08% (2,473)	4,103 % of Total: 100.00% (4,103)	70.24% Avg for View: 70.24% (0.00%)	1.55 Avg for View: 1.55 (0.00%)
1.	United States	0 (0.00%)	788 (31.84%)	1,185 (28.88%)	71.14%	1.59
2.	Spain	0 (0.00%)	481 (19.43%)	908 (22.13%)	75.44%	1.56
3.	United Kingdom	0 (0.00%)	251 (10.14%)	373 (9.09%)	75.34%	1.47
4.	France	0 (0.00%)	122 (4.93%)	347 (8.46%)	81.84%	1.25
5.	Russia	0 (0.00%)	30 (1.21%)	189 (4.61%)	20.63%	1.83
6.	Germany	0 (0.00%)	125 (5.05%)	175 (4.27%)	68.57%	1.62
7.	China	0 (0.00%)	73 (2.95%)	102 (2.49%)	70.59%	1.54
8.	Italy	0 (0.00%)	50 (2.02%)	78 (1.90%)	55.13%	1.85
9.	Canada	0 (0.00%)	58 (2.34%)	75 (1.83%)	85.33%	1.17
10.	Australia	0 (0.00%)	38 (1.54%)	59 (1.44%)	64.41%	1.71

Figure 4.1: PopFly analytics (A) Sessions per month. We observe an increase of the number of sessions (and users) over time; **(B) Sessions per country.** Note that most sessions (users) are from America, followed by Europe and Asia; **(C) Summary statistics of sessions per country.** With countries sorted based on the number of sessions.

sciences, operated by Cold Spring Harbor Laboratory, a not-for-profit research and educational institution.

In summary, the PopFly genome browser works not only as a database useful for retrieving population genetics information from this model species, but also as a research tool on which it is possible to perform population genomic studies and to test evolutionary hypotheses (*see 3.1.2. Testing evolutionary hypotheses using PopFly*). For these reasons, and based on both the impact of the previous genome browser developed by our group (PopDrowser), and the positive reception that PopFly already had from the community, we hope that this novel browser will become a reference tool in the *Drosophila* population genetics field.

What is next?

The current browser implementation has some limitations that should also be discussed. PopFly's major deficiency is that it uses the *D. melanogaster* reference version 5.57, because the DGN sequence data was mapped to this specific reference genome sequence. However, the *D. melanogaster* reference genome version 6 was released few years ago (Hoskins et al. 2015), and the newest population genomics studies in this species are carried out using this new reference. Thus, it seems clear that the DGN data pipeline and the PopFly browser itself should also shift to the new reference genome.

Interestingly, using the reference genome version 6 would allow the integration within PopFly of the huge amount of population genetics data obtained as part of the European Drosophila Population Genomics Consortium project (DrosEU; `http://droseu.net/`). This is a collaborative consortium of scientists and laboratories interested in evolutionary genetics and genomics of *Drosophila melanogaster*, founded in 2013. Its main objective is to cooperate closely in collecting, generating and analyzing genomic and environmental data for numerous *Drosophila* populations across Europe (and beyond). Our laboratory joined this consortium back at the end of 2016, and proposed to use PopFly as the project's main repository. This is currently and on-going project, but this process would require an effort and amount of time that could not be covered during the development of this work.

Additionally, another improvement that could not be performed yet but is scheduled to be completed in the future is the incorporation of multiple *Drosophila* species within the PopFly's framework, starting with the recently published *D. simulans* largest population genomics project (Signor et al. 2018). We think that the availability in the same software of a catalog of population genetics parameters estimated both in *D. melanogaster* and in this closely related species would facilitate and clarify testing more complex evolutionary hypotheses in *Drosophila*.

Related to the previous point, the development of PopFly provided us with the technological infrastructure and knowledge necessary to build up new genome browsers with data from other organisms. In this regard, and taking advantage of the huge amount of human genetic data gathered in the 1000 GP (Consortium et al. 2015), we decided to develop a novel human-based population genomics browser, named PopHuman (Casillas et al. 2018) (*see Appendix 7.2*).

This new browser allows the interactive visualization and retrieval of an extensive inventory of population genetics metrics estimated in non-overlapping windows along the chromosomes and in annotated genes for all 26 populations of the 1000GP (Casillas et al. 2018). The genome browser is open and freely accessible at `https://pophuman.uab.cat`; and it has been designed in order to work in the most commonly used web browsers (Chrome, Firefox, Safari, Microsoft Edge, Internet Explorer and Opera).

4.1.2 Integrative MKT software

A major issue in population genetics has been the accurate detection of the impact of natural selection along the genome. In this regard, many tests have been proposed, being the McDonald and Kreitman Test (Standard MKT, McDonald & Kreitman 1991) the most used method. Several modifications to the Standard MKT have been applied over the last years, leading to a battery of MKT-derived methods, with their own strengths and weaknesses (*see 1.2.3. Tests of selection*). However, even the improvements of both the quality and amount of sequencing data and the theoretical methods applied to such data, some major questions are still unsolved. First, what is the best method to quantify both the impact of positive and negative selection regimes acting on nucleotide variants? And then, once the methodological challenge has been solved, we should be able to address the true evolutionary questions: Which is the real rate of genome adaptation? Which fraction of the genome is under purifying selection?

Besides, the improvement of NGS technologies that lead to the sequencing of hundreds or even thousands of individuals from the same species provides us with the necessary data on which to empirically test evolutionary theoretical hypotheses. The availability of such datasets increases the request of developing high performance specific tools able to deal and analyze this information.

In order to face these challenges, we developed the Integrative MKT (iMKT). Note that iMKT accounts for two different concepts: (i) an extension of the asymptotic MKT which incorporates the estimators of the DGRP correction method (Mackay et al. 2012) to estimate several selection regimes; and (ii) the integration of five MKT derived methods within the same analysis framework: Standard MKT (McDonald & Kreitman 1991), FWW correction (Fay et al. 2001), DGRP correction (Mackay et al. 2012), asymptotic MK method (Messer & Petrov 2013, Haller & Messer 2017), and the new iMKT.

The performance, power and biological interpretation of the novel estimators of the fractions of negative selection developed as part of the iMKT method are discussed in a further section (*4.2.3. Prevalence of weakly deleterious selection and evidence of adaptive selection*). Thus, here the discussion is mainly focused on the technological development of the iMKT software.

The iMKT analysis framework has been implemented in two tools: an R package and a web-server. The first tool was completely developed as part of this thesis while the second one is out of this thesis' scope, and was developed by other members of our lab. Briefly, these services allow the user to apply the five MKT derived methods stated above and obtain results quickly and easily using R on any machine or through any web browser. Previously, we discussed and highlighted the importance of both PopFly and PopHuman genome browsers as central repositories of population genomics estimates for *D. melanogaster* and *Homo sapiens* model species, respectively. Thus, we decided to provide functionalities to directly link the iMKT software to both databases, which allows retrieving and analyzing population genomics information in a single step, demonstrating the utility of combining both tools.

The development of the iMKT software was inspired by two projects that also developed tools able to perform certain MKT methods: (i) the standard and generalized MKT website (Egea et al. 2008), and (ii) the asymptotic MK service (Haller & Messer 2017).

On the one hand, the standard and generalized MKT website (`http://mkt.uab.cat/mkt/`) is a web-server that allows performing MKTs using custom user data, not only for synonymous and non-synonymous changes, but also for other classes of regions and/or several loci. Briefly, the website has three different interfaces: (i) the standard MKT, where users can analyze several types of sites in a coding region, (ii) the advanced MKT, where users can compare two closely linked regions in the genome that can be either coding or non-coding, and (iii) the multi-locus MKT, where users can analyze many separate loci in a single multi-locus test (Egea et al. 2008). The input of this software is a set of nucleotide sequences together with their corresponding functional annotations, which are then processed in order to calculate the MKT table and the estimate of the rate of adaptive evolution. This represents a clear difference compared to iMKT input, which requires to previously process the data to obtain the number of polymorphic and divergent sites (see below). This feature made the standard and generalized MKT web-server very useful for dealing with sequence data directly retrieved from GenBank, but it has become inefficient for analyzing large datasets containing hundreds or even thousands of genomes.

On the other hand, the asymptotic MK web-server (`http://benhaller.com/messerlab/asymptoticMK.html`) provides an R-based implementation of the asymptotic MKT (Messer & Petrov 2013) as a web-based service, but it can also be run at the command line using *curl* (a tool for transferring data using various protocols), or as a local R script (Haller & Messer 2017). In brief, it allows performing the asymptotic MK method using also user custom polymorphism and divergence data.

The iMKT software incorporates most functionalities of the standard and generalized MKT web service (such as the possibility of using not only synonymous and non-synonymous changes but any classes of sites as neutral and putatively selected) and the complete framework of the asymptotic MKT, together with new developed functions, within a single unified tool. This allows the execution and comparison of the results obtained with several MKT-based methods on the same input data and computational environment, removing the possible bias caused by the application of different software with their own assumptions and limitations, and leading to comparable measures among the tests applied.

Input data and selection of the appropriate MK method

The initial data required to use the iMKT software consists of two tables (Figure 2.11). The first one includes the number of polymorphic sites (P) in each derived allele frequency (DAF) category for both neutral (0) and putatively selected (i) functional classes. The second table contains the number of divergent (D) and analyzed (m) sites for each class of sites. The format of input custom data is adapted from the one used by Haller & Messer (2017).

The number of polymorphic sites and DAF categories of the input data on which they are distributed determines the appropriate test to use. We observed that in order to obtain meaningful and unbiased estimates of α in the asymptotic methods, at least 10 DAF categories must be provided, but there is not any upper limit. Indeed, the performance of asymptotic methods depends on a balance between the number of DAF categories and the number of polymorphic sites in each of them. The higher both numbers are, the better the accuracy of these tests.

We are aware that the specificity of the input data format might be a limitation for some users. The estimation of divergent and polymorphic sites (along with their associated derived allele frequencies) is not a trivial process and requires the application of specific tools, which could be a limiting factor for using the iMKT software. To overcome this limitation, we designed and implemented a custom pipeline for analyzing the DGN and 1000GP data (`https://imkt.uab.cat/population-genetics-pipeline/`). The pipeline uses sequence alignments (in multiFASTA format) or variation data (in VCF) for a set of samples and an outgroup species and estimates the necessary nucleotide variation metrics to generate the iMKT input data. Currently, we are working on its scalability to any data source and its implementation as an on-line functionality in the iMKT web-server.

Then, we used empirical *D. melanogaster* and simulated genetic data (*see 3.3.1.i. Comparative analysis of four MKT methods at the single-gene level*) to determine the efficacy of each method in diverse situations based on the input data and the desired output. This allowed to design an analysis flowchart which helps deciding which is the optimal test to perform in every situation (Figure 4.2).

The first factor to consider is whether the input data consists of a single gene or a set of "concatenated" genes. The concept of concatenated genes refers to a set of genes that are analyzed as a whole, counting the number of polymorphic and divergent sites of all of them together. In general, asymptotic methods are not applicable on individual gene data. Haller & Messer (2017) demonstrated the greater power of the asymptotic MK test to estimate the true value of α, compared to the original non-asymptotic test. However, they also noticed the need for a large data set to obtain reasonably accurate results from the asymptotic test, showing that estimates of α from a single gene, or from a system with a very short divergence time, are unlikely to be meaningful. Our results also support these observations. Then, the second factor to take into account is whether to estimate the fractions of sites under negative selection or not. Only DGRP and iMKT methods are able to do so.

Within the tests recommended for single gene analysis, the DGRP correction should be used if there is interest in estimating the purifying selection fractions. If this is not the case, the decision depends on the distribution of fitness effects (DFE). If it is leptokurtic (L-shaped), either the standard MKT or the FWW method are recommended, because leptokurtic distributions have a smaller proportion of polymorphisms that are slightly deleterious (Eyre-Walker & Keightley 2007), and it is under this condition that the standard MKT and the FWW perform the best. Otherwise, the DGRP method should be used. Besides, for concatenated genes data, if there is interest in the negative selection fraction, the iMKT should be the first option. Otherwise, the asymptotic MKT is recommended. In the case that these methods do not work with the chosen DAF spectrum, the number of DAF bins used to classify polymorphic sites should be increased, if possible. Otherwise, the DGRP method should be used.

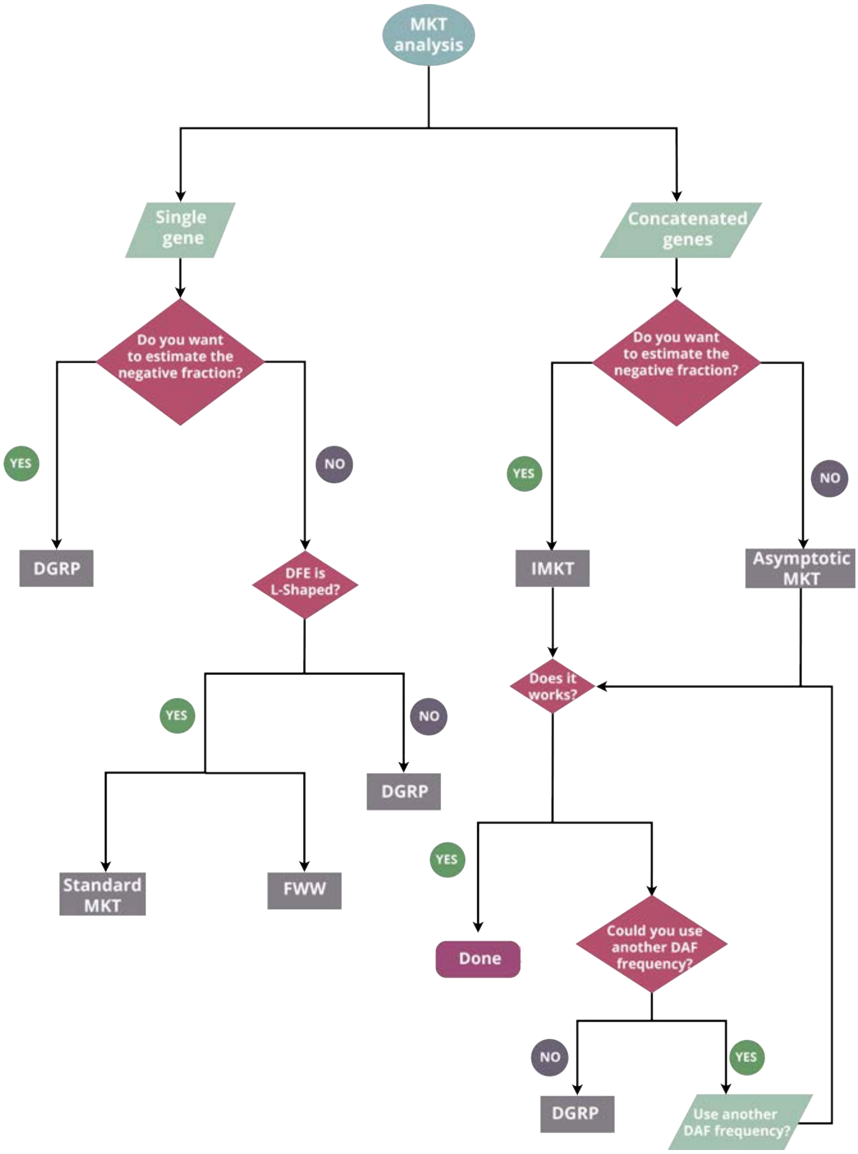


Figure 4.2: iMKT analysis flowchart. Flowchart shapes codes: circle shape is used to denote a begin. Parallelogram is used to represent input. The diamond represents a decision question. The squares represent the final decision. The lines with arrows determine the flow through the chart. (Figure designed by Jesús Murga).

The iMKT R package

The iMKT R package allows to compute the McDonald and Kreitman test on polymorphism and divergence genomic data provided by the user or automatically downloaded from PopFly (Hervas et al. 2017) or PopHuman (Casillas et al. 2018). It includes five MK derived methodologies, which allow inferring the rate of adaptive evolution (α) as well as the fraction of strongly deleterious (d), weakly deleterious (b), and neutral (f) sites.

The package is currently available at `https://github.com/BGD-UAB/iMKT`, it requires R environment version >= 3.3 to work, and its performance has been tested with successful results in the three major operating systems: Microsoft Windows, GNU/Linux and Macintosh OS.

The integration of MKT methods in the R framework environment allows performing fast and reproducible analyses (along with high-quality graphs) using custom user data. Moreover, it provides the possibility to be implemented within a larger work-flow allowing, at the same time, the integration and analysis of diverse layers of -omics information (ie. genomics, transcriptomics, epigenomics) in a unified environment. In addition, thanks to R's commitment to open-source software, developers can easily incorporate custom new methodologies into the iMKT framework, as well as contribute to the main source code (through GitHub's pull requests), allowing a continuous upgrade of the tool.

Along this work, we have specifically shown how to use the diverse functions implemented in the iMKT R package (*see 3.2. iMKT: an R package for the Integrative McDonald and Kreitman Test*) and we have also applied this software to perform a global comparison of the adaptation rate and negative selection fractions of six *D. melanogaster* meta-populations (*see 3.3.3. Detection of positive and purifying selection in the Drosophila genome*), demonstrating the utility of this novel tool for analyzing population genomics data. Indeed, most tables and graphs shown in these sections were directly retrieved from the R package output. The discussion of the estimates of adaptive and purifying selection obtained using the iMKT is addressed later in this work (*4.2.3 Prevalence of purifying selection and evidence of adaptive evolution*).

Altogether, we expect the iMKT R package to become a useful tool for scientists to analyze both their own custom data and two of the largest population genomics datasets available right now (DGN, Lack et al. 2015, 2016; 1000GP,Consortium et al. 2015).

What is next?

A major limitation of the iMKT R package right now is that it is not listed as an official R package in The Comprehensive R Archive Network (CRAN), which would reinforce its utility and diffusion while providing an easier way for installing it. Although the package fulfills the major CRAN requirements, the complete process for being accepted could not be completed during this thesis due to time restrictions. However, this action, together with the publication of the software and analyses performed, is scheduled to be completed in the near future.

In addition, as discussed previously, web services are the standard protocol for data exchange and communication. Thus, we also implemented all functionalities of the iMKT R package within a web-server to extend its availability. This work was mainly performed by other members of our group. The iMKT web server is open and freely accessible at site `https://imkt.uab.cat`. Its performance has been successfully tested in the most used web browsers (Chrome, Firefox, Safari, Microsoft Edge, Internet Explorer and Opera).

A major strength of this server compared to the R package on which it is based is that no programming (R) skills are required to obtain results in both the form of text and graphically (which be downloaded in a PDF file). In addition, it does not depend on any software besides a web browser. On the other hand, however, its performance is reduced compared to the R package because it acts mainly as a graphical user interface that executes in the back-end the R functions from the iMKT package. Thus, some time is wasted in these steps of data traffic and exchange. Additionally, the amount of machine memory that can be used is limited by the server in which it is stored, while the R package is only limited by the machine on which it is executed.

The iMKT web-server home page shows direct links to the following four sections:

- MKT custom analysis: upload custom polymorphism and divergence data files, select the preferred test(s) and get the results.

- PopFly data analysis: allows selecting genes from the PopFly database or upload a custom list, then select the population(s) of interest and perform the desired MK test(s).

- PopHuman data analysis: allows selecting genes from the PopHuman database or upload a custom list, then select the population(s) of interest and perform the desired MK test(s).

- Learn more (About): menu with four entries, which are (i) iMKT R package: explanation and link to download the iMKT R package to perform the analysis in the user's local R environment; (ii) *Drosophila* and human data: description of the *Drosophila* and human data available at iMKT; (iii) Population genetics pipeline: summary of the pipeline developed and possibility to download it; (iv) MK methodologies: brief description about how the different MK-derived methodologies work (Standard MKT, FWW method, DGRP method, Asymptotic MKT, iMKT).

In addition, it also has a contact section that allows users to make suggestions and share any issues or questions related to this web server with the development and maintenance team.

Once published, we expect the iMKT web-server to have a great reception from the scientific community and become a useful tool due to its simplicity to use and easy availability.

4.2 Population genomics of *Drosophila melanogaster*

In the last years, several population genomics studies in *D. melanogaster* have been performed (Langley et al. 2012, Mackay et al. 2012, Pool et al. 2012, Huang et al. 2014, Grenier et al. 2015, Kao et al. 2015). These studies included specific analyses of nucleotide variation and adaptation patterns along the genome, and the assessment of the impact of the major genomic determinants, like recombination or gene expression, in the establishment of such patterns; using single or few populations.

However, each of these studies used its own pipeline for processing genome data and for estimating the inventory of population genetics parameters necessary to test evolutionary hypotheses, making the comparison of results from diverse projects very though. To overcome this challenge, the DGN project compiled all available population genomic sequencing reads and re-assembled them against a single common reference genome assembly, using a unified pipeline (Lack et al. 2015). The DGN consortium performed a very preliminary analysis of the data, mainly oriented to clarify the genetic structure of such populations (Lack et al. 2016), but a global population genomics analysis of the data was still missing.

This thesis faces the challenge: (i) estimating a complete inventory of population genetics parameters in all available *D. melanogaster* populations, and (ii) providing a global description of nucleotide variation patterns, as well as their major determinants, all around the globe. Even that most analyses are not new in terms of originality (e.g., estimating genome-wide patterns of polymorphism or the rate of adaptive evolution in diverse chromosome arms are routine steps in population genomics studies), this work is indeed original and new in the sense that it provides a comprehensive overview and comparison of estimates computed using a common approach. Specifically, metrics are calculated in 30 *D. melanogaster* world-wide populations and the analyses were performed in six *Drosophila* meta-populations.

Nevertheless, we are aware that the analyses presented in this thesis have some limitations that do not allow us to draw concise conclusions. First, our main goal here is to provide a general view of the evolutionary dynamics of genetic variation. Thus, there is still a need of performing additional and more specific analyses that could not be covered during the development of this project due to time constraints. Second, the DGN sequence data is primarily aimed at SNP-oriented analysis, and aside from inversion calling and the detection of short insertions/deletions (indels), structural variation is not addressed (Lack et al. 2016).

Nucleotide variation estimates obtained in this work are somewhat higher than previous values. There are two possible explanations for it. First, we are analyzing aggregations of populations, i.e., meta-populations, which are composed by samples from diverse populations. Thus, genetic differences between the populations may cause an increase in the estimates of nucleotide diversity for the resulting meta-populations. Second, these differences could be partially explained by the methodological approach that we used. We filtered genome sequences to account for heterozygosity (and "pseudo-heterozygosity"), admixture and identity by descent (IBD), re-coding the affected genome regions as "N" and thus, removing these genome tracks from the analyses. These regions are usually expected to show lower levels of nucleotide diversity and hence, by excluding them, the overall estimates are increased. Finally, we used a subset of samples for each population to estimate the battery of population genetics parameters, considering only informative positions (i.e., non-"N"), and ensuring a constant sample size for all analyzed windows.

4.2.1 Impact of demography on populations structure

The demographic history of a population leaves a footprint in the patterns of polymorphism, which could mimic the one produced by natural selection. For example, a reduction of nucleotide diversity could be caused either by a population bottleneck or a selective sweep event (*see 1.2.2. Genome-wide signatures of selection*). Thus, understanding the demographic history of a species is an important requirement to be able to make inferences about how natural selection and other evolutionary forces have shaped the genome variation landscape.

D. melanogaster is a cosmopolitan species that originated from sub-Saharan Africa. The ancestral Afro-tropical population suffered a population bottleneck followed by a population expansion about 60 kya, which promoted the fixation of several beneficial mutations (Stephan & Li 2007, Singh et al. 2012). In addition, the ancestral population colonized Europe and North America around 19,000 and 200 years ago, respectively, leaving signatures in the genome of local adaptation (Duchen et al. 2012).

To understand the structure of populations in this species, assess whether or not this structure is consistent with the species known demographic history, and justify the analysis of samples at the meta-population level, we performed a phylogenetic tree reconstruction using F_{ST} values from Lack et al. (2016) (Figure 3.12). Even that F_{ST} metrics used here are restricted to inversion-free chromosome arms, Lack et al. (2015) demonstrated only small and non-statistically significant effects of inversions on genetic differentiation levels.

Geographic structure is apparent, especially between African (ancestral) and non-African (colonizer) populations, with the latter group showing a common reduced gene pool apparently resulting from past population bottlenecks (Grenier et al. 2015, Lack et al. 2016).

Regarding the African cluster, F_{ST} estimates revealed patterns similar to those of Pool et al. (2012) for these populations. In detail, population differentiation was particularly low among southern African populations and somewhat elevated among Ethiopian samples, which showed moderate differentiation from other sub-Saharan samples probably due to a past bottleneck (Pool et al. 2012, Lack et al. 2015).

In the non-African cluster, populations are structured into several groups. First, the three populations from America are genetically very close, with the western population (USW) being more differentiated to the eastern ones (RAL, USI), a pattern previously found (Caracristi & Schlötterer 2003, Campo et al. 2013). The prevailing demographic model for *D. melanogaster* suggests that the colonization of North America took place very recently with Europe as the source of the founder flies (David & Capy 1988). This model implies a rapid demographic growth involving both population and range expansion from eastern to western North America. In addition, recent studies demonstrated that North American populations have mainly European but partly African ancestry (Kao et al. 2015, Pool 2015, Bergland et al. 2016), which is also supported by our phylogenetic reconstruction (Figure 3.12). The Australia population is genetically closer to the American than the Asian populations, consistent with Reinhardt et al. (2014) observations.

The two European populations (FR, NTH) are also grouped together and show low differentiation among them, as expected (Grenier et al. 2015). Surprisingly, the Egypt population (EG) falls within the non-African cluster, even though it is an African population. This could be explained by the Sahara desert acting as a natural barrier which separates both African regions, and causing higher genetic differentiation between North and South Africa populations than between North Africa and European populations. Thus, we considered EG together with FR and NTH populations in a single Europe/North Africa meta-population, as suggested by Lack et al. (2016).

Finally, some studies have shown that Asian lines are the most differentiated samples (Schlotterer et al. 2005, Grenier et al. 2015), and that additional population bottlenecks may have impacted the China population (Laurent et al. 2011). Our results are in agreement with an independent migration from Africa to Asia (David et al. 1976, Grenier et al. 2015), as the China population is closer to the other non-African samples than to the African ones.

4.2.2 Genome-wide nucleotide variation and recombination patterns

Polymorphism and divergence

Genome-wide molecular population analyses showed that patterns of polymorphism (π) differ: (i) among different autosomal regions, and (ii) between the X chromosome and the autosomes. Polymorphism is lower in the centromeric and telomeric regions of the autosomes, and this reduction is gradual and spans several kilobases, while in chromosome X there is only a small decrease of variability in the telomeric region. These patterns are shared among all populations from this species. On the contrary, the X chromosome has an overall reduced level of polymorphism in non-African populations, but not in the African ones. This strong reduction in diversity was previously reported (Begun & Aquadro 1993, Baudry et al. 2004, Mackay et al. 2012, Langley et al. 2012, Huang et al. 2014, Grenier et al. 2015) and presumably results from the bottleneck that occurred during the expansion out of sub-Saharan Africa.

We observe that levels of polymorphism are in agreement with the populations structure discussed in the previous section. The Asian population shows the lowest levels of π (0.0426), consistent with previous results (Grenier et al. 2015 found $\pi \sim 0.03$; and Lack et al. 2016 $\pi = 0.0401$) and with the particular demographic history of this population, affected by multiple bottlenecks (Laurent et al. 2011).

Then, when comparing the American and European populations, we observe that American samples show somewhat higher diversity because of their partial African ancestry (Kao et al. 2015, Pool 2015, Bergland et al. 2016). Indeed, we found a polymorphism level of $\pi = 0.0539$, similar to Mackay et al. (2012) ($\pi = 0.0531$) and Lack et al. (2016) (π in the range 0.0511-0.0569). However, the nucleotide diversity levels estimated for the Europe/North African meta-population are higher in this work ($\pi = 0.0535$) than in Lack et al. (2016) (π in the range 0.0466-0.0495). This is caused by the aggregation of the Egypt population together with the European ones, resulting in higher estimates of polymorphism (see above). The Oceania population shows similar nucleotide diversity levels to Europe/N. Africa and America, with an average of $\pi = 0.0529$ (Lack et al. 2016, $\pi = 0.0516$, Grenier et al. 2015, $\pi \sim 0.04$). Finally, within the African populations, genomic diversity is higher in the South African meta-population ($\pi = 0.0813$) than in the Equatorial African one ($\pi = 0.0757$), as previously described by Pool et al. (2012), Lack et al. (2015, 2016).

Besides, nucleotide divergence (k) patterns show high peaks of divergence in the centromeric regions shared by all meta-populations, and observed using either *D. simulans* or *D. yakuba* as outgroup species, in agreement with the results reported by Mackay et al. (2012), Langley et al. (2012), Huang et al. (2014). There are at least three non-exclusive reasons that could explain it: (i) a reduced quality of alignments in these regions producing more spurious polymorphisms, (ii) higher mutation rates in those regions, or (iii) higher fixation of slightly deleterious mutations due to low recombination reducing the efficacy of selection (Ràmia 2015). The latter possibility is discussed in following sections. Regarding k between *D. melanogaster* and the outgroup species *D. simulans* and *D. yakuba*, meta-populations analyzed in this work showed levels of k in the ranges $0.0633 - 0.0698$ and $0.128 - 0.136$, respectively. Differences among meta-populations were not statistically significant, as expected. The estimated rates are in concordance with the ones obtained for the North American population analyzed in the Freeze 2.0 of the DGRP (Huang et al. 2014), where k was estimated to be 0.062 and 0.1283 using *D. simulans* and *D. yakuba* outgroup species, respectively.

Recombination

Recombination rate appears to be one major determinant of the levels of nucleotide variation and adaptation along the genome (*discussed in following sections*). Thus, a fine-scale recombination map is a prerequisite of studies seeking to estimate the influence of natural selection on the genome. Although the inferences of recombination and selection used in this work rely on the same data and have the potential to distort each other, Chan et al. (2012) demonstrated that their method is robust to the influence of positive selection, and that in general, ρ estimates show good agreement with existing experimental estimates of recombination, such as the one developed by Comeron et al. (2012).

The ρ metrics estimated here are also consistent with the recombination rates in cM/Mb by Comeron et al. (2012). Genome-wide patterns are similar, with a decrease of recombination rates near the centromeric and telomeric regions of autosomes and at both extremes of the X chromosome (Figure 3.17). Hitchhiking and background selection (Charlesworth 1994) in these large regions of reduced recombination may also contribute to this relative reduction in values of ρ, as noticed by Langley et al. (2012).

In addition, ρ and recombination rates in cM/Mb show a positive and significant correlation for all meta-populations analyzed (Figure 4.3). This association is observed in all chromosome arms independently, with correlation coefficients ranging from 0.4 to 0.75 (all with an associated $p < 10^{-11}$). For the X chromosome, the correlations coefficients are: 0.39 for Asia; 0.48 for Oceania; 0.68 for America; 0.69 for Europe/N. Africa; 0.72 for Equatorial Africa; and 0.7 for Southern Africa (all $p < 10^{-09}$). However, as we are comparing the historical recombination rate ($\rho = 4N_e r$) with recombination rates in cM/Mb, the correlation coefficients that we obtain are somewhat moderate.

These results are also consistent with the previously ρ recombination map developed by Langley et al. (2012) as part of the Drosophila Population Genomics Project 1 (DPGP1). Specifically, authors estimated ρ using 37 lines of the RAL population and the LDhat version 2.1 software package (McVean et al. 2004). The LDhat software was the first robust LD-based recombination rate calculator, aimed mainly to estimate ρ using human genetic data. Thus, it has certain limitations for dealing with high-diversity genetic datasets such as the ones of *D. melanogaster* (*see Chan et al. (2012) paper about their new tool LDhelmet, based on LDhat, for a deep discussion and comparison among both tools*).

We also find extensive fine-scale variation across all chromosomes and meta-populations. The most outstanding difference is the higher overall recombination rate in African populations compared to non-African ones, consistent with Chan et al. (2012) results. Because ρ is estimated as the composite parameter of r and N_e, where N_e is the effective population size and r is the (female) rate of recombination per generation, this difference might be explained (at least partially) by differences in the effective population size of such data sets (Chan et al. 2012).

Besides, we observed that recombination rates in the X chromosome are higher than in the autosomes in all populations except Asia. In detail, we observe that ρ estimates for the Asian population are not consistent with the estimates obtained for the other populations, and thus, these results must be interpreted with caution. It is possible that the Asian population actually has lower recombination rates due to its reduced effective population size and the multiple bottlenecks experienced by this population. However, we found that the Asian genome sequences have, on average, lower quality (i.e., they show a larger fraction of non-informative nucleotides) and extensive genome

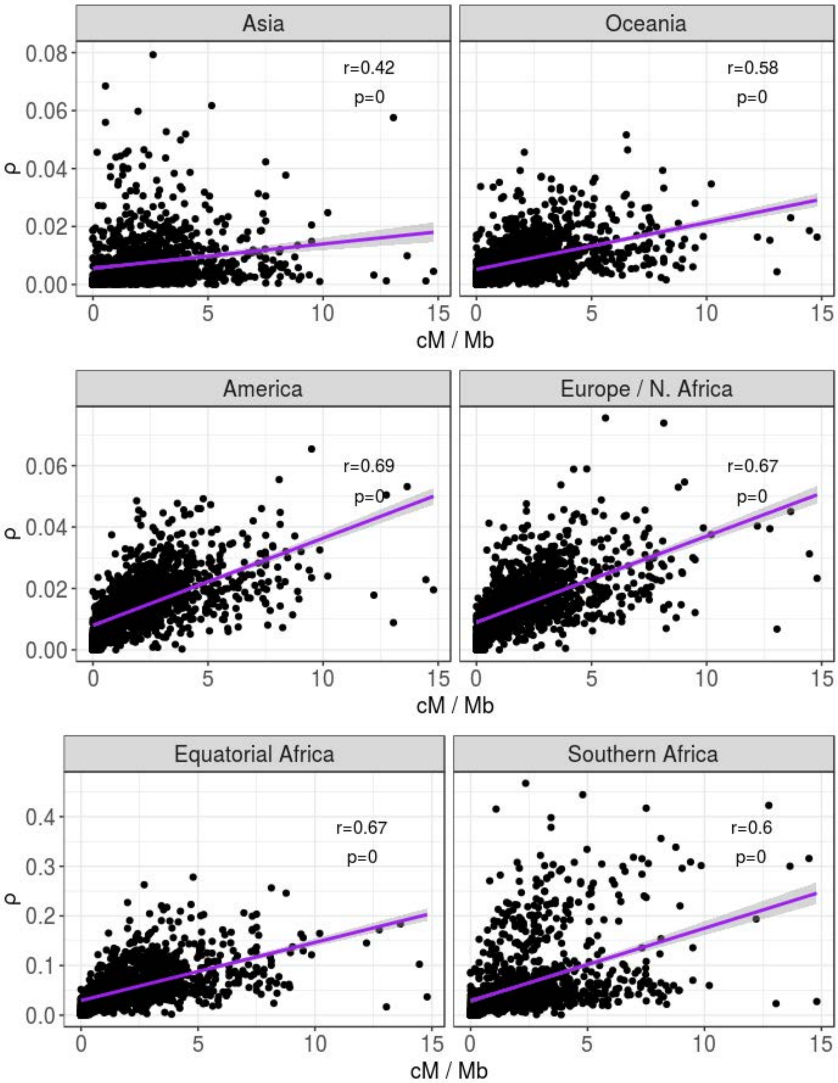


Figure 4.3: Comparison between experimental and computational recombination maps. Spearman's rank correlation coefficients together with their associated p-values. The experimental high-resolution map estimates correspond to Comeron et al. (2012). Metrics and correlation are for 100 kb non-overlapping windows covering the complete euchromatic genome. Note that Y axis scales differ between rows of the plot, but in all cases we observe a positive and statistically significant correlation.

regions masked by IBD (Lack et al. 2016), which might bias our estimates of ρ.

The higher ρ values in the X chromosome compared to the autosomes are much more pronounced in African than non-African populations (Figure 3.18). There are three possible explanations for the differences observed between populations, as discussed by Chan et al. (2012). First, the historical population bottleneck experienced by non-African populations as part of the out-of-Africa process. The effect of a population bottleneck on LD is stronger on the X chromosome than on the autosomes due to its reduced effective population size (Wall et al. 2002). Furthermore, Chan et al. (2012) have shown using simulations that bottlenecks which reduce the N_e tend to cause the LDhelmet method to underestimate the true recombination rate. Second, the impact of polymorphic inversions may be greater in African populations, since they have a higher frequency of polymorphic inversions in the autosomes and in the X chromosome. Thus, the increase of African X chromosome recombination rate could be partially attributed to the effect of polymorphic inversions disturbing the normal patterns of recombination. Third, as *D. melanogaster* males are hemizygous for the X chromosome, deleterious mutations are more exposed to the action of natural selection, leading to a more efficient role of selection on the X chromosome compared to autosomes. This last possibility is discussed later in this work (*4.2.5. The faster-X hypothesis*).

So in summary, our ρ estimates appear to be consistent with both experimental (Comeron et al. 2012) and LD-based computational (Langley et al. 2012, Chan et al. 2012) recombination rate maps previously developed. In addition, they are also in agreement with the demographic and genetic (i.e., polymorphic inversions frequencies) histories of the diverse populations. Overall, we can conclude that our high-resolution ρ maps may also reflect the pattern of recombination (r) per base pair on even finer resolutions. However, it remains possible that hitchhiking or other forces may interfere with local estimates of ρ.

4.2.3 Prevalence of purifying selection and evidence of adaptive evolution

A main observation from population genomics analyses is that adaptive and purifying selection is pervasive in the genomes of most studied species, especially in those with a high N_e such as *Drosophila*. Briefly, deleterious mutations arise continuously and a large fraction of segregating sites are undergoing weak deleterious selection, while adaptive selection is also ubiquitous (Casillas & Barbadilla 2017).

The strength of the diverse selection regimens shaping the patterns of molecular evolutionary change can be estimated by the comparison of polymorphism and divergence rates between different classes of sites. These classes of sites are either assumed to be neutral, or putatively selected (i.e., the target of selection).

Mutations in 4-fold sites have been tipically considered as a proxy for neutrality since long (Kimura 1968) and used in major population genomics studies (Mackay et al. 2012), and so we did in this work. However, there are a couple of reasons why this type of sites might not be completely neutral. First, some 4-fold degenerated sites may be in linkage disequilibrium with non-synonymous strongly selected sites, thus behaving, in the practice, as being selected. Second, analyses of the codon usage bias suggested that synonymous mutations could also be subject to selection (Hershberg & Petrov 2009). In this regard, Lawrie et al. (2013) estimated that over 20% of mutations in 4-fold positions in *D. melanogaster* are indeed deleterious and thus, destined to rapidly disappear. Short intron sites have been shown to be evolving (nearly) neutrally in *Drosophila* (Parsch et al. 2010) and can be used as an alternative proxy for neutrality. Short intron sites are defined as those sites falling in introns of less than length 86bp, 16bp away from the intron start and 6bp away from the intron end in order to eliminate any functional sequences at the edges of the introns (Haddrill et al. 2005, Lawrie et al. 2013).

In this regard, both the already published McDonald and Kreitman-based tests (McDonald & Kreitman 1991, Fay et al. 2001, Mackay et al. 2012, Messer & Petrov 2013) and the extended asymptotic method (integrative MKT) applied along this work have the flexibility to consider any class of sites as putatively selected and neutral.

Estimators of the Integrative MKT and the fractions of sites under purifying selection

Mackay et al. (2012) proposed three novel estimators of the purifying selection fractions (d: strongly deleterious, b: weakly deleterious, and f: neutral) as part of the extended MKT method applied in that project (i.e., DGRP correction). Briefly, the DGRP method relies on a pre-established derived allele frequency (DAF) cut-off used to differentiate between neutral and weakly deleterious sites within the putatively selected class. If this cut-off is not well defined, it may cause an under- or over-estimation of the number of sites in any of the two categories, leading to biased α estimates (Campos et al. 2014).

To overcome this limitation, we implemented these estimators together with the asymptotic MKT method (Messer & Petrov 2013, Haller & Messer 2017) in the iMKT (*see 2.4.1.v iMKT approach, and 3.2.1 Estimators of the integrative MKT*). The iMKT does not consider any prior frequency cut-off to trim polymorphisms, and relies on the fit of the data to an asymptotic model. Hence, the cut-off is defined by the data itself, which facilitates the analyses of diverse types of data (i.e., human -vs- *Drosophila* or autosomes -vs- X chromosome data sets).

Using the DGRP data, Mackay et al. (2012) found that averaged over the entire genome, 58.5% of the segregating sites are neutral or nearly neutral, 1.9% are weakly deleterious, and 39.6% are strongly deleterious; although they also found great variability between the X chromosome and autosomes, chromosome regions and classes of sites. Regarding non-synonymous sites, these are the most constrained ($d = 77.6\%$), the fraction of weakly deleterious sites is much larger than the genome average ($b = 3.8\%$) and consequently, the neutral fraction is reduced ($f = 18.6\%$). Later, Castellano (2016) re-analyzed the DGRP data considering a different genomic scale (using the complete genome as a single window) and found, for non-synonymous positions: (i) a larger fraction of strongly deleterious sites ($d = 81\%$), (ii) similar proportion of weakly deleterious sites ($b = 3\%$), and (iii) a slightly lower fraction of neutral sites ($f = 15\%$), explained by the increase of d, because $f = 1 - (d + b)$).

Here, we analyzed non-synonymous sites (0-fold degenerated coding positions) as putatively selected class, and we applied both DGRP and iMKT methods to all six meta-populations at diverse genomic scales (genome, autosomes, each autosomic chromosome arm, and X chromosome) using the higher quality assembly of *D. simulans* (Hu et al. 2013) as outgroup species (*see 3.3.3. Detection of positive and purifying selection in the Drosophila genome*).

Our estimates for the six *Drosophila* meta-populations (Table 3.13) are consistent with the estimates reported by Mackay et al. (2012) and (Castellano 2016), but we found a larger fraction of strongly deleterious sites ($d \sim 86\%$) than these studies. We also observe that the fraction of weakly deleterious sites estimated using the DGRP correction ($b \sim 2.5\%$) is much lower than using the iMKT ($b \sim 7.5\%$); and the opposite for the fraction of neutral sites ($f_{DGRP} \sim 11.5\%$; $f_{iMKT} \sim 6.5\%$). As discussed previously, it is possible that we are under-estimating the fraction of b with the DGRP method because of the DAF cut-off used to distinguish between deleterious and neutral segregating mutations. A considerable proportion of mutations with slightly deleterious fitness effects may reach frequencies above this cut-off because they are in linkage disequilibrium with adaptive mutations, and thus, they will be wrongly assumed to be functionally neutral.

However, we also found an association between metrics estimates and the sample size of each meta-population, using both the DGRP and the iMKT methods. Specifically, populations with lowest sample sizes (Asia, $n = 18$ and Oceania, $n = 15$) show the highest d and lowest b values, followed by the three meta-populations composed of 30 individuals from three populations each (America, Europe/N. Africa and Southern Africa), and finally, the meta-population with the largest sample size (Equatorial Africa, $n = 50$) presents the lowest d and highest b values. As here we are comparing samples with different n, results must be interpreted with caution. Instead, further investigation should be performed (using equilibrated sample sizes, ideally) in order to get more reliable and accurate results. In addition, these novel estimators need to be tested against simulations with known output to assess their accuracy and error rates.

Even though this is an initial general analysis, our results allow us to discuss some general trends. First, the pressure of natural selection against non-synonymous mutations with strongly negative fitness effects seems to be stronger in the short past (i.e., in the *D. melanogaster* - *D. simulans* lineage compared with *D. yakuba*). Specifically, more than 80% of new 0-fold coding mutations have strongly deleterious effects. Then, we found that the proportion of weakly deleterious mutations within the non-synonymous class is much higher than previously though (~7.5% of sites), a pattern maintained in all meta-populations and chromosomes analyzed (Table 3.13). Consequently, we detected a smaller fraction of non-synonymous sites which are effectively neutral (~6.5%).

In summary, the genome of *D. melanogaster* appears to be under hard selective constraint, with mutations having strongly deleterious effects being removed quickly from the genetic pool. However, some mutations with weakly deleterious effects are able to escape from natural selection pressures because they are linked to advantageous mutations or located in regions of very low recombination where selection is less efficient.

Adaptive evolution

Initial studies using population genomics data in *D. melanogaster* such as the DGRP (Mackay et al. 2012) showed that on average 25.2% of the fixed sites between *D. melanogaster* and *D. yakuba* are adaptive, ranging from 30% in introns, to 24% for non-synonymous sites, to 7% in UTR sites.

However, Messer & Petrov (2013) demonstrated few years later that these values were under-estimated because they were obtained using a non-asymptotic approach that was unable to remove accurately a large fraction of weakly deleterious mutations. Indeed, they found that the rate of adaptive evolution in *D. melanogaster*, using *D. simulans* as outgroup, is as high as 0.57 (with a confidence interval in the range 0.54-0.6). This study, performed using 130 lines of the RAL population (from the DGRP) also estimated the original α from the standard MKT (McDonald & Kreitman 1991), which is 0.407, thus demonstrating the power of their novel method.

Here, using *D. simulans* as outgroup and the iMKT (which incorporates the asymptotic method of Messer & Petrov 2013), we found somewhat higher α values, ranging on average from 0.6 to 0.65 for non-African populations and from 0.68 to 0.69 for African ones. In detail, adaptation rates for African populations are higher in all chromosome arms (Table 3.12); and values in the X chromosome are higher than in the autosomes in all samples, consistent with previous observations (Mackay et al. 2012, Langley et al. 2012). Overall, our results promote us to state that a large fraction of non-synonymous fixations in *D. melanogaster* are adaptive, which demonstrates the influence of recurrent and strong positive selection pressures shaping the genome variation patterns in such species.

However, there is a major question in evolutionary biology (and population genomics) which still needs to be solved: *What is the adaptive significance of this amount of positive selection?* Even though we usually refer to positive and adaptive selection as synonymous terms, they might not be. If weakly deleterious mutations are constantly fixed in the genome (like in regions under Hill-Robertson interference, *see the following section*), the opportunity for compensatory mutations to also reach fixation increases. We consider compensatory mutations those that restore the negative effect of the previously fixed deleterious mutations. Thus, it is possible that many of those non-synonymous variants fixed by positive selection are not adaptive (they do not provide an innovative new feature to the organism) but only compensatory (Kimura 1985).

In order to obtain a complete picture of how adaptation occurs, population genomics data will need to be integrated with other phenotypic multi-omics layers of information such as transcriptomics or epigenomics, leading toward a population -omics synthesis era (Casillas & Barbadilla 2017).

4.2.4 Genomic determinants of the adaptation rate and the HRi

The levels of nucleotide diversity and rates of adaptation are known to vary along the genome (Mackay et al. 2012). There are many genomic determinants which influence the establishment of these distinctive patterns, such as gene density and function, and differences in mutation or recombination rates.

The first observation from population genomics studies is that recombination rate shows a positive correlation with the polymorphism level for every analyzed variant (Begun & Aquadro 1992, Mackay et al. 2012, Huang et al. 2014). Mackay et al. (2012) hypothesized that the correlation between polymorphism and recombination is due to the reduced N_e in regions of low recombination. This would imply that selection is also reduced on such regions, and these should show a reduced fraction of strongly deleterious and positively selected sites. Our results support this hypothesis, as we observe that, in all meta-populations, the fraction of strongly deleterious sites and the adaptation rate increase as recombination does so (Figure 4.4). Besides, the proportions of weakly deleterious and effectively neutral sites show the opposite pattern.

Because at any time there are genetic variants in LD simultaneously selected in the genome, these variants interfere with each other, reducing the efficacy of natural selection and inducing a cost of linkage known as the HRi (Hill & Robertson 1968). As recombination increases, the linkage disequilibrium between alleles is reduced, variants can segregate more freely, and consequently, the efficacy of natural selection is increased (*see 1.2.1. Determinants of patterns of genetic variation*).

In this regard, the genome can be divided in two distinctive types of regions depending on their linkage disequilibrium: non-linked selected blocks (NLSB), and linked selection blocks (LSB) (Barrón 2015). In the first class of regions, the neutral theory can be applied as a null model for the interpretation of genetic variation (Lewontin & Krakauer 1973), and they they constitute ~40% of the *D. melanogaster* genome (27% in autosomes, 77% in the X chromosome, Barrón 2015). Besides, in LSD (regions of high linkage and low recombination), HRi is predicted to constantly occur, which means that ~60% of the genome, especially in the autosomes, seems to be in a sub-optimal situation regarding natural selection efficacy. The neutral theory is not longer valid as a null model in these regions, which means that the development of new baseline models of genome variation (such as the one developed by Corbett-Detig et al. 2015, combining BGS and hitchhiking and incorporating polymorphism, recombination rate, and density of functional elements in the genome) are required to assess the impact of selection on this type of regions.

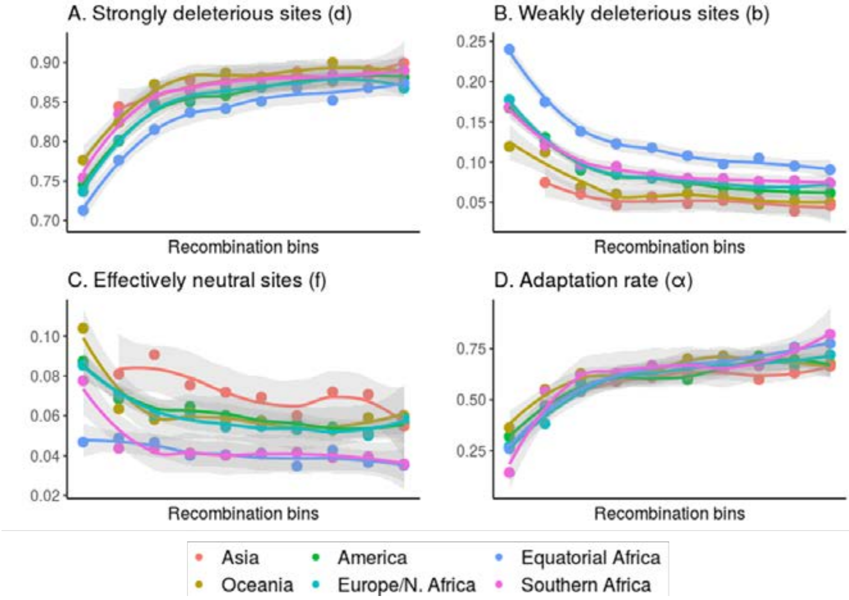


Figure 4.4: The impact of recombination on natural selection regimes. Average estimates of **(A)** strongly deleterious sites, **(B)** weakly deleterious sites, **(C)** effectively neutral sites and, **(D)** adaptation rate; computed in ten bins of recombination, for all six meta-populations. The complete set of 13,753 genes were ranked according to the gene-associated ρ values in each population and grouped in ten bins (each bin containing 1,375 genes). Note that bin 1 of the Asia population is not shown because we could not estimate b and thus the other metrics were biased. Lines and confidence intervals (shaded in grey) were computed using logical regression methods (loess curve). The values in the X axis do not correspond to the estimate of the rate of recombination (ρ). Instead, they correspond to each of the ten bins.

Castellano et al. (2016) provided an approach to quantify the adaptive potential of a genome. By estimating the optimal recombination threshold from which mutations segregate independently and from which adaptation is not restricted by HRi (r_{opt}), it is then possible to estimate the cost of linkage (L_{HRi}). Specifically, authors found that r_{opt} was 2 cM/Mb, and that HRi reduces the evolutionary adaptation rate of the *D. melanogaster* genome by an average of 27% (using the RAL North American population, $n = 205$).

Here, we quantified this effect in diverse wild-derived populations with different demographic and migratory histories in order to characterize which are the HRi load dynamics operating in *Drosophila*, and their impact on the molecular evolutionary rate. We found that adaptation is substantially impeded by HRi in all meta-populations, with L_{HRi} estimates ranging from $\sim 19.5\%$ to $\sim 24\%$ (Table 3.14), values similar to the estimate provided by Castellano et al. (2016). Thus, our results reinforce the importance of HRi as a major force operating at the species level, but with a weak strength in local adaptation processes.

Besides recombination, another genomic determinant of the adaptation rate is gene density (Hey & Kliman 2002). Genes located in gene-poor regions are expected to show higher rates of adaptation, whereas genes in gene-rich regions would show lower rates (Castellano 2016). Assuming that the mutation and the recombination rates are constant and uniform along the genome, new mutations arising in gene-rich regions are more prone to be in linkage disequilibrium with other variants, reducing the efficacy of selection via HRi; while mutations appearing in gene-poor regions are expected to be more effectively targeted by natural selection. Indeed, Castellano et al. (2016) found a negative correlation between the rate of adaptive evolution and the density of selected sites across the genome (Spearman's correlation coefficient = -0.69, with and associated $p < 0.001$). This correlation was independent from gene functions (correlation coefficient = -0.75, $p < 0.001$ excluding immune and testes specific genes).

However, our results can not confirm Castellano et al. (2016) observations, as we do not obtain any correlation coefficient between gene density and the rate of adaptation with statistical support (Figure 3.22). Increasing the number of bins does not lead to significant results (e.g., correlation coefficient = -0.17, $p > 0.3$ using the American meta-population and genes grouped in 30 bins). Hence, further analyses are required in order to elucidate whether Castellano et al. (2016) observations of the negative association among gene density and the rate of adaptive evolution can be replicated or not, and its implications in the evolutionary dynamics of *D. melanogaster*.

4.2.5 The faster-X hypothesis

Although the *Drosophila* X chromosome is usually similar to the autosomes in size and cytogenetic appearance, theoretical models predict that its hemizygosity in males may cause unusual patterns of accelerated evolution (Vicoso & Charlesworth 2006). This is known as the faster-X hypothesis (Charlesworth et al. 1987).

In an equilibrium population with equal numbers of males and females, equal X-linked and autosomal mutation rates, and no natural selection, the expected ratio of X-linked vs. autosomal diversity is 3/4. However, in the ancestral (African) populations of *D. melanogaster* studied here, we observed X-linked diversity levels (corrected by the X/A effective population size factor) higher than the autosomal average, consistent with previous results (Hutter et al. 2007, Langley et al. 2012). In addition, the X chromosome shows increased levels of nucleotide divergence compared to autosomes. These observations could be explained by a higher mutation rate in the X. However, Keightley et al. (2009) demonstrated that there are no statistical differences between the mutation rates in autosomes and chromosome X.

Another plausible explanation for the observed X polymorphisms levels is the distinctive effect of background selection in autosomes and the X chromosome (Charlesworth 1994). Because *Drosophila* males are hemizigous for the X chromosome (XO), deleterious mutations arising in it are more exposed to natural selection and thus more efficiently purged from the gene pool, compared to autosomes. However, background selection has not been predicted to strongly influence diversity in regions of the *Drosophila* genome subject to moderate or high rates of recombination (Langley et al. 2012) and thus, this process alone can not explain the observed differences.

Besides, Vicoso & Charlesworth (2006) proposed that differences in recombination rates (generally higher on the X) could account for observed deviations of the X/A diversity ratio. Because the autosomes show lower recombination rates than the X chromosome, they are more affected by linked selection (and HRi) and thus they show a greater diversity reduction. Here, we found higher X recombination rates (ρ) for all meta-populations except Asia (discussed above). The inflated X-linked ρ metrics were much more pronounced in the African populations than in non-African ones. In terms of linked and non-linked selection blocks, Barrón (2015) showed that the larger rate of recombination in the X with respect to the autosomes makes that 50% more sites are selectively independent in this chromosome than in the autosomes. In addition, we also found evidences of higher adaptation rates in the X chromosome than autosomes for all analyzed meta-populations, with up to an increase of $\sim$20% of fixed adaptive variants in that specific chromosome. Regarding negative selection fractions we can observe a slight increase of strongly deleterious sites and a clear decrease of weakly deleterious variants in all populations (except Asia).

Relative to the African populations, the non-African ones showed an X/A diversity ratio significantly lower than 1, even after correcting for the X/A effective population size factor (4/3). This might be explained by the demographic histories of such populations, as it is known that demographic events such as population bottlenecks or founder events may lead to a disproportionate reduction in X-linked diversity (Wall et al. 2002). However, demographic models such as the one proposed by Pool & Nielsen (2008) cannot completely explain the X-linked and autosomal variation in the non-African sample.

An alternative explanation for the stronger reduction of diversity in the non-African X chromosomes is a stronger hitchhiking effect on the X chromosome relative to autosomes for those populations. This is consistent with our observations that the X/A recombination ratio is much lower in non-African than in African populations.

Overall, all molecular population genomic analyses performed along this work support the faster-X hypothesis. The X chromosome shows increased rates of polymorphism for African populations (in non-African populations we observe a reduction of π in the X chromosome compared to the autosomes, probably caused by demographic events), faster rates of molecular evolution, a higher fraction of adaptive and strongly deleterious variants, and a lower proportion of weakly deleterious sites, relative to the autosomes.

Overall, the general population genomics overview presented in this thesis allows to get a global picture of the patterns of nucleotide variation and adaptation along the *Drosophila* genome. By analyzing six meta-populations from all around the globe (Asia, Oceania, America, Europe/North Africa, Equatorial Africa and Southern Africa), we provide a comprehensive description and a unified vision of the population genomics dynamics in this model species.

We found that most patterns are shared by all populations, such as the distribution of polymorphism, divergence and historical recombination values along the chromosomes, or the accelerated rate of evolution in the X chromosome. The impact of both positive and purifying selection, as well as the adaptive constraint caused by HRi, are also long term processes that affect all populations, i.e., they are orthogonal to the species actual geographic distribution.

Besides, we also confirmed local differences among populations subject to distinctive demographic and environmental pressures. In general, African populations have higher nucleotide diversity levels, recombination and adaptation rates than non-African ones, consistent with the demographic history of such species.

In conclusion, this thesis should serve as a reference starting point for future and more detailed population genomics analysis using the DGN resource data.

5. Conclusions

5. Conclusions

General conclusions

- We developed two novel bioinformatics tools to facilitate the visualization and analysis of population genomics data: the population genomics web-browser PopFly, and the R statistical package iMKT.

- We applied both tools to perform an initial population genomics analysis in *D. melanogaster* using the DGN data, to get a global picture of the nucleotide variation and adaptation patterns along the genome, and assess the relative impact of the genomic determinants of genetic variation in six meta-populations spanning five continents.

PopFly, the *Drosophila* population genomics browser

- PopFly contains the broadest catalog of population genetics estimates in *D. melanogaster*, storing more than 4,000 tracks with information of almost one thousand samples from 30 world-wide natural populations from the *Drosophila Genome Nexus* project data.

- The user-friendly graphical web interface of PopFly allows the visualization and retrieval of functional annotations, estimates of nucleotide diversity and divergence, linkage disequilibrium statistics, recombination rate metrics, a battery of neutrality tests, and population differentiation parameters, at different window sizes through the euchromatic chromosomes.

- In addition, we have also developed and implemented new utilities and support resources within the PopFly framework to facilitate performing population genetics analyses on-the-fly and retrieving sequence data.

- The development of PopFly also provided us with the technological infrastructure and knowledge necessary to build up genome browsers in other species, such as PopHuman.

iMKT: an R package for the Integrative McDonald and Kreitman Test

- The Integrative McDonald and Kreitman Test (iMKT) package for the R software environment allows computing the McDonald and Kreitman test on polymorphism and divergence genomic data provided by the user or automatically downloaded from PopFly or PopHuman.

- It calculates five MKT derived tests (Standard MKT, MKT with the FWW correction, MKT with the DGRP correction, asymptotic MK method, and the novel iMKT). All methods allow inferring the rate of adaptive evolution (α). In addition, the MKT with the DGRP correction and the integrative MKT allow inferring the fraction of strongly deleterious (d), weakly deleterious (b), and neutral (f) sites.

- The iMKT R package has also been implemented, outside the scope of this thesis, in a freely accessible web-server which allows the user to calculate the five MKT derived methods stated above and perform analyses quickly and easily through any web browser.

Population genomics of *Drosophila melanogaster*

- Polymorphism levels are reduced near the centromeres and telomeres of autosomes and at both extremes of the X chromosome. They are also lower in non-African populations ($\pi = 0.005$) than in African ones ($\pi = 0.008$); and the X chromosome has an overall reduced level of diversity in non-African populations.

- Divergence (relative to both *D. simulans* and *D. yakuba* outgroup species) is extremely high in the centromeric regions of all chromosomes, and increased in the X chromosome compared to autosomes.

- Historical recombination rate estimates are also reduced near the centromeres and telomeres. ρ is higher in African populations compared to non-African ones, and in the X chromosome relative to the autosomes.

- The genome of *D. melanogaster* is under hard selective constraint, with more than 85% of new 0-fold coding mutations having strongly deleterious effects, $\sim 7.5\%$ being weakly deleterious and $\sim 6.5\%$ being effectively neutral.

- A large fraction of non-synonymous fixations between *D. melanogaster* and *D. simulans* are adaptive ($\alpha = 0.66$). The X chromosome shows an increased rate of adaptive evolution relative to autosomes ($\alpha_X = 0.78$).

- Recombination is positively correlated with both the fraction of adaptive fixations and the fraction of strongly deleterious sites, while it is negatively correlated with the fraction of weakly deleterious sites.

- Overall, the adaptation rate in *Drosophila* is substantially impeded by Hill-Robertson interference, which diminishes the rate of adaptive evolution by approximately $\sim 22\%$.

6. References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F. et al. (2000), 'The genome sequence of drosophila melanogaster', *Science* **287**(5461), 2185–2195.

Andolfatto, P. (2005), 'Adaptive evolution of non-coding dna in drosophila', *Nature* **437**(7062), 1149.

Aoyagi, N. & Wassarman, D. A. (2000), 'Genes encoding drosophila melanogaster rna polymerase ii general transcription factors: diversity in tfiia and tfiid components contributes to gene-specific transcriptional regulation', *The Journal of Cell Biology* **150**(2), F45–F50.

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E. M., Bowden, R., Aneas, I., Broxholme, J. et al. (2012), 'A fine-scale chimpanzee genetic map from population sequencing', *science* **336**(6078), 193–198.

Avery, O. T., MacLeod, C. M. & McCarty, M. (1944), 'Studies on the chemical nature of the substance inducing transformation of pneumococcal types', *Journal of Experimental Medicine* **79**(2), 137–158.

Ayala, F. J., Tracey, M. L., Barr, L. G., McDonald, J. F. & Pérez-Salas, S. (1974), 'Genetic variation in natural populations of five drosophila species and the hypothesis of the selective neutrality of protein polymorphisms', *Genetics* **77**(2), 343–384.

Barrón, M. G. (2015), Patrones de variación nucleotídica y cartografía de bloques de selección ligada en el genoma de Drosophila melanogaster, PhD thesis, Universitat Autònoma de Barcelona.

Baudry, E., Viginier, B. & Veuille, M. (2004), 'Non-african populations of drosophila melanogaster have a unique origin', *Molecular biology and evolution* **21**(8), 1482–1491.

Begun, D. J. & Aquadro, C. F. (1992), 'Levels of naturally occurring dna polymorphism correlate with recombination rates in d. melanogaster', *Nature* **356**(6369), 519–520.

Begun, D. J. & Aquadro, C. F. (1993), 'African and north american populations of drosophila melanogaster are very different at the dna level', *Nature* **365**(6446), 548.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N. et al. (2007), 'Population genomics: whole-genome analysis of polymorphism and divergence in drosophila simulans', *PLoS biology* **5**(11), e310.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. (2013), 'Genbank', *Nucleic Acids Research* **41**(D1), D36–D42.

Bergland, A. O., Tobler, R., Gonzalez, J., Schmidt, P. & Petrov, D. (2016), 'Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in drosophila melanogaster', *Molecular ecology* **25**(5), 1157–1174.

Bergman, C. M. & Haddrill, P. R. (2015), 'Strain-specific and pooled genome sequences for populations of drosophila melanogaster from three continents.', *F1000Research* **4**.

Berry, A. J., Ajioka, J. & Kreitman, M. (1991), 'Lack of polymorphism on the drosophila fourth chromosome resulting from selection', *Genetics* **129**(4), 1111–1117.

Bhérer, C., Campbell, C. L. & Auton, A. (2017), 'Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales', *Nature communications* **8**, 14994.

Birky, C. W. & Walsh, J. B. (1988), 'Effects of linkage on rates of molecular evolution', *Proceedings of the National Academy of Sciences* **85**(17), 6414–6418.

Boc, A., Diallo, A. B. & Makarenkov, V. (2012), 'T-rex: a web server for inferring, validating and visualizing phylogenetic trees and networks', *Nucleic Acids Research* **40**(W1), W573–W579.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. (1995), 'The hitchhiking effect on the site frequency spectrum of dna polymorphisms.', *Genetics* **140**(2), 783–796.

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., Stein, L. et al. (2016), 'Jbrowse: a dynamic web platform for genome visualization and analysis', *Genome biology* **17**(1), 66.

Campo, D., Lehmann, K., Fjeldsted, C., Souaiaia, T., Kao, J. & Nuzhdin, S. (2013), 'Whole-genome sequencing of two north american drosophila melanogaster populations reveals genetic differentiation and positive selection', *Molecular ecology* **22**(20), 5084–5097.

Campos, J. L., Halligan, D. L., Haddrill, P. R. & Charlesworth, B. (2014), 'The relation between recombination rate and patterns of molecular evolution and variation in drosophila melanogaster', *Molecular biology and evolution* **31**(4), 1010–1028.

Caracristi, G. & Schlötterer, C. (2003), 'Genetic differentiation between american and european drosophila melanogaster populations could be attributed to admixture of african alleles', *Molecular biology and evolution* **20**(5), 792–799.

Casillas, S. (2008), Development and application of bioinformatic tools for the representation and analysis of genetic diversity, PhD thesis, Universitat Autònoma de Barcelona.

Casillas, S. & Barbadilla, A. (2006), 'Pda v. 2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous dna', *Nucleic acids research* **34**(suppl_2), W632–W634.

Casillas, S. & Barbadilla, A. (2017), 'Molecular population genetics', *Genetics* **205**(3), 1003–1035.

Casillas, S., Mulet, R., Villegas-Mirón, P., Hervas, S., Sanz, E., Velasco, D., Bertranpetit, J., Laayouni, H. & Barbadilla, A. (2018), 'Pophuman: the human population genomics browser', *Nucleic Acids Research* **46**(D1), D1003–D1010.

Castellano, D. (2016), Estimación de la huella de la selección natural y el efecto Hill-Robertson a lo largo del genoma de Drosophila melanogaster, PhD thesis, Universitat Autònoma de Barcelona.

Castellano, D., Coronado-Zamora, M., Campos, J. L., Barbadilla, A. & Eyre-Walker, A. (2016), 'Adaptive evolution is substantially impeded by hill–robertson interference in drosophila', *Molecular biology and evolution* **33**(2), 442–455.

Chan, A. H., Jenkins, P. A. & Song, Y. S. (2012), 'Genome-wide fine-scale recombination rate variation in drosophila melanogaster', *PLOS Genetics* **8**(12), 1–28.

Chang, W. (2012), *R Graphics Cookbook: Practical Recipes for Visualizing Data*, O'Reilly Media, Inc.

Charlesworth, B. (1994), 'The effect of background selection against deleterious mutations on weakly selected, linked variants', *Genetics Research* **63**(3), 213–227.

Charlesworth, B. (2010), *Elements of evolutionary genetics*, Roberts Publishers.

Charlesworth, B., Coyne, J. & Barton, N. (1987), 'The relative rates of evolution of sex chromosomes and autosomes', *The American Naturalist* **130**(1), 113–146.

Charlesworth, B., Morgan, M. & Charlesworth, D. (1993), 'The effect of deleterious mutations on neutral molecular variation.', *Genetics* **134**(4), 1289–1303.

Charlesworth, B., Nordborg, M. & Charlesworth, D. (1997), 'The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations', *Genetics Research* **70**(2), 155–174.

Charlesworth, D., Charlesworth, B. & Morgan, M. (1995), 'The pattern of neutral molecular variation under the background selection model.', *Genetics* **141**(4), 1619–1632.

Chen, H., Patterson, N. & Reich, D. (2010), 'Population differentiation as a test for selective sweeps', *Genome research* **20**(3), 393–402.

Chen, X., Hiller, M., Sancak, Y. & Fuller, M. T. (2005), 'Tissue-specific tafs counteract polycomb to turn on terminal differentiation', *Science* **310**(5749), 869–872.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M. et al. (2007), 'Evolution of genes and genomes on the drosophila phylogeny', *Nature* **450**, 203 – 218.

Comeron, J. M., Ratnappan, R. & Bailin, S. (2012), 'The many landscapes of recombination in drosophila melanogaster', *PLoS genetics* **8**(10), e1002905.

Comeron, J. M., Williford, A. & Kliman, R. (2008), 'The hill–robertson effect: evolutionary consequences of weak selection and linkage in finite populations', *Heredity* **100**(1), 19.

Consortium, T. . G. P. et al. (2015), 'A global reference for human genetic variation', *Nature* **526**(7571), 68.

Coop, G., Wen, X., Ober, C., Pritchard, J. K. & Przeworski, M. (2008), 'High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans', *science* **319**(5868), 1395–1398.

Coordinators, N. R. (2016), 'Database resources of the national center for biotechnology information', *Nucleic acids research* **44**(Database issue), D7.

Corbett-Detig, R. B. & Hartl, D. L. (2012), 'Population genomics of inversion polymorphisms in drosophila melanogaster', *PLOS Genetics* **8**(12), 1–15.

Corbett-Detig, R. B., Hartl, D. L. & Sackton, T. B. (2015), 'Natural selection constrains neutral diversity across a wide range of species', *PLoS Biology* **13**(4), e1002112.

Da Lage, J.-L., Thomas, G. W., Bonneau, M. & Courtier-Orgogozo, V. (2018), 'Evolution of salivary glue genes in drosophila species', *bioRxiv* p. 359190.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T. et al. (2011), 'The variant call format and vcftools', *Bioinformatics* **27**(15), 2156–2158.

Darwin, C. (1859), *On the origin of species by means of natural selection*, Murray, London.

David, J., Bocquet, C. & Pla, E. (1976), 'New results on the genetic characteristics of the far east race of drosophila melanogaster', *Genetics Research* **28**(3), 253–260.

David, J. R. & Capy, P. (1988), 'Genetic variation of drosophila melanogaster natural populations', *Trends in Genetics* **4**(4), 106–111.

De Castro, S., Peronnet, F., Gilles, J.-F., Mouchel-Vielh, E. & Gibert, J.-M. (2018), 'bric à brac (bab), a central player in the gene regulatory network that mediates thermal plasticity of pigmentation in drosophila melanogaster', *PLoS genetics* **14**(8), e1007573.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M. et al. (2011), 'A framework for variation discovery and genotyping using next-generation dna sequencing data', *Nature genetics* **43**(5), 491.

Dobzhansky, T. (1937), 'Genetics and the origin of species', *Columbia University Press, New York, USA* .

Dobzhansky, T. (1970), 'Genetics of the evolutionary process', *Columbia University Press, New York* .

Dobzhansky, T. & Sturtevant, A. H. (1938), 'Inversions in the chromosomes of drosophila pseudoobscura', *Genetics* **23**(1), 28–64.

Duchen, P., Živković, D., Hutter, S., Stephan, W. & Laurent, S. (2012), 'Demographic inference reveals african and european admixture in the north american drosophila melanogaster population', *Genetics* pp. genetics–112.

Egea, R., Casillas, S. & Barbadilla, A. (2008), 'Standard and generalized mcdonald–kreitman test: a website to detect selection by comparing different classes of dna sites', *Nucleic acids research* **36**(suppl_2), W157–W162.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R. & Ashburner, M. (2005), 'The sequence ontology: a tool for the unification of genome annotations', *Genome biology* **6**(5), R44.

Ellegren, H. (2014), 'Genome sequencing and population genomics in non-model organisms', *Trends in ecology & evolution* **29**(1), 51–63.

Endler, J. (1986), 'Natural selection in the wild', *Princeton University Press, New Jersey, USA* .

Eyre-Walker, A. (2002), 'Changing effective population size and the mcdonald-kreitman test', *Genetics* **162**(4), 2017–2024.

Eyre-Walker, A. & Keightley, P. D. (2007), 'The distribution of fitness effects of new mutations', *Nature Reviews Genetics* **8**(8), 610–618.

Fay, J. C. & Wu, C.-I. (2000), 'Hitchhiking under positive darwinian selection', *Genetics* **155**(3), 1405–1413.

Fay, J. C., Wyckoff, G. J. & Wu, C.-I. (2001), 'Positive and negative selection on the human genome', *Genetics* **158**(3), 1227–1234.

Fisher, R. A. (1922), 'On the interpretation of χ 2 from contingency tables, and the calculation of p', *Journal of the Royal Statistical Society* **85**(1), 87–94.

Fisher, R. A. (1930), 'The distribution of gene ratios for rare mutations', *Proc. R. Soc. Edinb* **50**, 205–222.

Fiston-Lavier, A.-S., Singh, N. D., Lipatov, M. & Petrov, D. A. (2010), 'Drosophila melanogaster recombination rate calculator', *Gene* **463**(1), 18–20.

Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S. et al. (2009), 'Ensembl's 10th year', *Nucleic acids research* **38**(suppl_1), D557–D562.

Ford, E. B. (1971), Ecological genetics, *in* 'Ecological genetics, Ed.3', Springer, pp. 1–11.

Franssen, S. U., Nolte, V., Tobler, R. & Schlötterer, C. (2015), 'Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental drosophila melanogaster populations', *Molecular biology and evolution* **32**(2), 495–509.

Fu, Y.-X. (1997), 'Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection', *Genetics* **147**(2), 915–925.

Fu, Y.-X. & Li, W.-H. (1993), 'Statistical tests of neutrality of mutations', *Genetics* **133**(3), 693–709.

Furey, T. S. (2006), 'Comparison of human (and other) genome browsers', *Human genomics* **2**(4), 266.

Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. (2015), 'Recent selective sweeps in north american drosophila melanogaster show signatures of soft sweeps', *PLoS genetics* **11**(2), e1005004.

Garud, N. R. & Petrov, D. A. (2016), 'Elevation of linkage disequilibrium above neutral expectations in ancestral and derived populations of drosophila melanogaster', *Genetics* pp. genetics–115.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. et al. (2004), 'Bioconductor: open software development for computational biology and bioinformatics', *Genome biology* **5**(10), R80.

Gramates, L. S., Marygold, S. J., Santos, G. d., Urbano, J.-M., Antonazzo, G., Matthews, B. B., Rey, A. J., Tabone, C. J., Crosby, M. A., Emmert, D. B. et al. (2016), 'Flybase at 25: looking to the future', *Nucleic acids research* p. gkw1016.

Grenier, J. K., Arguello, J. R., Moreira, M. C., Gottipati, S., Mohammed, J., Hackett, S. R., Boughton, R., Greenberg, A. J. & Clark, A. G. (2015), 'Global diversity lines–a five-continent reference panel of sequenced drosophila melanogaster strains', *G3: Genes, Genomes, Genetics* **5**(4), 593–603.

Guillén, Y., Rius, N., Delprat, A., Williford, A., Muyas, F., Puig, M., Casillas, S., Ràmia, M., Egea, R., Negre, B. et al. (2014), 'Genomics of ecological adaptation in cactophilic drosophila', *Genome biology and evolution* **7**(1), 349–366.

Haddrill, P. R., Charlesworth, B., Halligan, D. L. & Andolfatto, P. (2005), 'Patterns of intron sequence evolution in drosophila are dependent upon length and gc content', *Genome biology* **6**(8), R67.

Haerty, W., Jagadeeshan, S., Kulathinal, R. J., Wong, A., Ram, K. R., Sirot, L. K., Levesque, L., Artieri, C. G., Wolfner, M. F., Civetta, A. et al. (2007), 'Evolution in the fast lane: rapidly evolving sex-related genes in drosophila', *Genetics* **177**(3), 1321–1335.

Haldane, J. (1932), 'The causes of evolution', *Princeton University Press, Princeton, NJ* .

Hales, K. G., Korey, C. A., Larracuente, A. M. & Roberts, D. M. (2015), 'Genetics on the fly: a primer on the drosophila model system', *Genetics* **201**(3), 815–842.

Haller, B. C. & Messer, P. W. (2017), 'asymptoticmk: A web-based tool for the asymptotic mcdonald–kreitman test', *G3: Genes, Genomes, Genetics* **7**(5), 1569–1575.

Hardy, G. H. (1908), 'Mendelian proportions in a mixed population', *Science* **28**(706), 49–50.

Harris, H. (1966), 'Enzyme polymorphisms in man', *Proc. R. Soc. Lond. B* **164**(995), 298–310.

Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W. J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R. et al. (2009), 'Wormbase: a comprehensive resource for nematode research', *Nucleic acids research* **38**(suppl_1), D463–D467.

Hershberg, R. & Petrov, D. A. (2009), 'General rules for optimal codon choice', *PLoS Genetics* **5**(7), e1000556.

Hervas, S., Sanz, E., Casillas, S., Pool, J. E. & Barbadilla, A. (2017), 'Popfly: the drosophila population genomics browser', *Bioinformatics* **33**(17), 2779–2780.

Hey, J. & Kliman, R. M. (2002), 'Interactions between natural selection, recombination and gene density in the genes of drosophila', *Genetics* **160**(2), 595–608.

Hill, W. G. & Robertson, A. (1966), 'The effect of linkage on limits to artificial selection', *Genetics Research* **8**(3), 269–294.

Hill, W. & Robertson, A. (1968), 'Linkage disequilibrium in finite populations', *Theoretical and Applied Genetics* **38**(6), 226–231.

Hiller, M., Chen, X., Pringle, M. J., Suchorolski, M., Sancak, Y., Viswanathan, S., Bolival, B., Lin, T.-Y., Marino, S. & Fuller, M. T. (2004), 'Testis-specific taf homologs collaborate to control a tissue-specific transcription program', *Development* **131**(21), 5297–5308.

Hippel, E. v. & Krogh, G. v. (2003), 'Open source software and the "private-collective" innovation model: Issues for organization science', *Organization science* **14**(2), 209–223.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian journal of statistics* pp. 65–70.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., Svirskas, R. et al. (2015), 'The release 6 reference sequence of the drosophila melanogaster genome', *Genome research* pp. gr–185579.

Hoskins, R. A., Smith, C. D., Carlson, J. W., Carvalho, A. B., Halpern, A., Kaminker, J. S., Kennedy, C., Mungall, C. J., Sullivan, B. A., Sutton, G. G. et al. (2002), 'Heterochromatic sequences in a drosophila whole-genome shotgun assembly', *Genome biology* **3**(12), research0085–1.

Howe, D. G., Bradford, Y. M., Conlin, T., Eagle, A. E., Fashena, D., Frazer, K., Knight, J., Mani, P., Martin, R., Moxon, S. A. T. et al. (2012), 'Zfin, the zebrafish model organism database: increased support for mutants and transgenics', *Nucleic acids research* **41**(D1), D854–D860.

Hu, T. T., Eisen, M. B., Thornton, K. R. & Andolfatto, P. (2013), 'A second-generation assembly of the drosophila simulans genome provides new insights into patterns of lineage-specific divergence', *Genome Research* **23**(1), 89–98.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A. M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R. F. et al. (2014), 'Natural variation in genome architecture among 205 drosophila melanogaster genetic reference panel lines', *Genome research* **24**(7), 1193–1208.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T. et al. (2002), 'The ensembl genome database project', *Nucleic acids research* **30**(1), 38–41.

Hudson, R. R. (1987), 'Estimating the recombination parameter of a finite population model without selection', *Genetics Research* **50**(3), 245–250.

Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J. & Ayala, F. J. (1994), 'Evidence for positive selection in the superoxide dismutase (sod) region of drosophila melanogaster.', *Genetics* **136**(4), 1329–1340.

Hudson, R. R., Kreitman, M. & Aguadé, M. (1987), 'A test of neutral molecular evolution based on nucleotide data', *Genetics* **116**(1), 153–159.

Hudson, R. R., Slatkin, M. & Maddison, W. (1992), 'Estimation of levels of gene flow from dna sequence data', *Genetics* **132**(2), 583–589.

Hutter, S., Li, H., Beisswanger, S., De Lorenzo, D. & Stephan, W. (2007), 'Distinctly different sex ratios in african and european populations of drosophila melanogaster inferred from chromosome-wide snp data', *Genetics* .

Hutter, S., Vilella, A. J. & Rozas, J. (2006), 'Genome-wide dna polymorphism analyses using variscan', *BMC Bioinformatics* **7**(1), 409.

Ihaka, R. & Gentleman, R. (1996), 'R: a language for data analysis and graphics', *Journal of computational and graphical statistics* **5**(3), 299–314.

James, J. E., Piganeau, G. & Eyre-Walker, A. (2016), 'The rate of adaptive evolution in animal mitochondria', *Molecular ecology* **25**(1), 67–78.

Jukes, T. & Cantor, C. (1969), 'Evolution of protein molecules', *Mammalian Protein Metabolism. Academic Press, New York* pp. 21–123.

Kao, J. Y., Zubair, A., Salomon, M. P., Nuzhdin, S. V. & Campo, D. (2015), 'Population genomic analysis uncovers african and european admixture in drosophila melanogaster populations from the south-eastern united states and caribbean islands', *Molecular ecology* **24**(7), 1499–1509.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J. et al. (2003), 'The ucsc genome browser database', *Nucleic acids research* **31**(1), 51–54.

Keenan, K., McGinnity, P., Cross, T. F., Crozier, W. W. & Prodöhl, P. A. (2013), 'diversity: an r package for the estimation and exploration of population genetics parameters and their associated errors', *Methods in Ecology and Evolution* **4**(8), 782–788.

Keightley, P. D. & Eyre-Walker, A. (2010), 'What can we learn about the distribution of fitness effects of new mutations from dna sequence data?', *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**(1544), 1187–1193.

Keightley, P. D. & Eyre-Walker, A. (2012), 'Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small', *Journal of Molecular Evolution* **74**(1), 61–68.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S. & Blaxter, M. (2009), 'Analysis of the genome sequences of three drosophila melanogaster spontaneous mutation accumulation lines', *Genome research* pp. gr–091231.

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. (2010), 'Bigwig and bigbed: enabling browsing of large distributed datasets', *Bioinformatics* **26**(17), 2204–2207.

Kimura, M. (1955), 'Stochastic processes and distribution of gene frequencies under natural selection', *Cold Spring Harbor Symposia on Quantitative Biology* **20**, 33–53.

Kimura, M. (1957), 'Some problems of stochastic processes in genetics', *The Annals of Mathematical Statistics* pp. 882–901.

Kimura, M. (1968), 'Evolutionary rate at the molecular level', *Nature* **217**(5129), 624–626.

Kimura, M. (1980), 'A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences', *Journal of molecular evolution* **16**(2), 111–120.

Kimura, M. (1983), *The neutral theory of molecular evolution*, Cambridge University Press.

Kimura, M. (1985), 'The role of compensatory neutral mutations in molecular evolution', *Journal of Genetics* **64**(1), 7.

Kirkpatrick, M. (2010), 'How and why chromosome inversions evolve', *PLoS biology* **8**(9), e1000501.

Koh, T.-W., He, Z., Gorur-Shandilya, S., Menuz, K., Larter, N. K., Stewart, S. & Carlson, J. R. (2014), 'The drosophila ir20a clade of ionotropic receptors are candidate taste and pheromone receptors', *Neuron* **83**(4), 850–865.

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T. et al. (2010), 'Fine-scale recombination rate differences between sexes, populations and individuals', *Nature* **467**(7319), 1099.

Kreitman, M. (1983), 'Nucleotide polymorphism at the alcohol dehydrogenase locus of drosophila melanogaster', *Nature* **304**(5925), 412.

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., Langley, C. H. & Pool, J. E. (2015), 'The drosophila genome nexus: A population genomic resource of 623 drosophila melanogaster genomes, including 197 from a single ancestral range population', *Genetics* **199**(4), 1229–1241.

Lack, J. B., Lange, J. D., Tang, A. D., Corbett-Detig, R. B. & Pool, J. E. (2016), 'A thousand fly genomes: An expanded drosophila genome nexus', *Molecular Biology and Evolution* **33**(12), 3308–3313.

Lanfear, R., Kokko, H. & Eyre-Walker, A. (2014), 'Population size and the rate of evolution', *Trends in ecology & evolution* **29**(1), 33–41.

Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C. G., Schrider, D. R., Pool, J. E., Langley, S. A., Suarez, C., Corbett-Detig, R. B., Kolaczkowski, B. et al. (2012), 'Genomic variation in natural populations of drosophila melanogaster', *Genetics* pp. genetics–112.

Laurent, S. J., Werzner, A., Excoffier, L. & Stephan, W. (2011), 'Approximate bayesian analysis of drosophila melanogaster polymorphism data reveals a recent colonization of southeast asia', *Molecular biology and evolution* **28**(7), 2041–2051.

Lawrie, D. S., Messer, P. W., Hershberg, R. & Petrov, D. A. (2013), 'Strong purifying selection at synonymous sites in d. melanogaster', *PLoS genetics* **9**(5), e1003527.

Lercher, M. J. & Hurst, L. D. (2002), 'Human snp variability and mutation rate are higher in regions of high recombination', *Trends in genetics* **18**(7), 337–340.

Lewontin, R. (1964), 'The interaction of selection and linkage. i. general considerations; heterotic models', *Genetics* **49**(1), 49–67.

Lewontin, R. (1974), 'The genetic basis of evolutionary change', *Columbia University Press, New York, USA* .

Lewontin, R. C. (1970), 'The units of selection', *Annual Review of Ecology and Systematics* **1**(1), 1–18.

Lewontin, R. C. & Hubby, J. L. (1966), 'A molecular approach to the study of genic heterozygosity in natural populations. ii. amount of variation and degree of heterozygosity in natural populations of drosophila pseudoobscura', *Genetics* **54**(2), 595–609.

Lewontin, R. & Kojima, K.-i. (1960), 'The evolutionary dynamics of complex polymorphisms', *Evolution* **14**(4), 458–472.

Lewontin, R. & Krakauer, J. (1973), 'Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms', *Genetics* **74**(1), 175–195.

Li, H. & Durbin, R. (2010), 'Fast and accurate long-read alignment with burrows–wheeler transform', *Bioinformatics* **26**(5), 589–595.

Li, Q., Zhou, Y., Jiao, Y., Zhang, Z., Bai, L., Tong, L., Yang, X., Sommer, B., Hofestädt, R. & Chen, M. (2016), 'Dato: an atlas of biological databases and tools', *Journal of integrative bioinformatics* **13**(2), 30–38.

Li, W.-H., Wu, C.-I. & Luo, C.-C. (1985), 'A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes.', *Molecular biology and evolution* **2**(2), 150–174.

Lunter, G. & Goodson, M. (2011), 'Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads', *Genome research* **21**(6), 936–939.

Lynch, M. & Walsh, B. (2007), *The origins of genome architecture*, Vol. 98, Sinauer Associates Sunderland (MA).

Mackay, T. F., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M. et al. (2012), 'The drosophila melanogaster genetic reference panel', *Nature* **482**(7384), 173.

Mair, P., Hofmann, E., Gruber, K., Hatzinger, R., Zeileis, A. & Hornik, K. (2015), 'Motivation, values, and work design as drivers of participation in the r open source project for statistical computing', *Proceedings of the National Academy of Sciences* **112**(48), 14788–14792.

Matute, D. R., Gavin-Smyth, J. & Liu, G. (2014), 'Variable post-zygotic isolation in drosophila melanogaster/d. simulans hybrids', *Journal of evolutionary biology* **27**(8), 1691–1705.

Maxam, A. M. & Gilbert, W. (1977), 'A new method for sequencing dna', *Proceedings of the National Academy of Sciences* **74**(2), 560–564.

McDonald, J. H. & Kreitman, M. (1991), 'Adaptive protein evolution at the adh locus in drosophila', *Nature* **351**(6328), 652.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. & Donnelly, P. (2004), 'The fine-scale structure of recombination rate variation in the human genome', *Science* **304**(5670), 581–584.

Messer, P. W. & Petrov, D. A. (2013), 'Frequent adaptation and the mcdonald–kreitman test', *Proceedings of the National Academy of Sciences* **110**(21), 8615–8620.

Metcalf, C. E. & Wassarman, D. A. (2007), 'Nucleolar colocalization of taf1 and testis-specific tafs during drosophila spermatogenesis', *Developmental Dynamics* **236**(10), 2836–2843.

Misra, S., Crosby, M. A., Mungall, C. J., Matthews, B. B., Campbell, K. S., Hradecky, P., Huang, Y., Kaminker, J. S., Millburn, G. H., Prochnik, S. E. et al. (2002), 'Annotation of the drosophila melanogaster euchromatic genome: a systematic review', *Genome biology* **3**(12), research0083–1.

Morgan, T. H. (1915), *The mechanism of Mendelian heredity*, Holt.

Muller, H. J. (1927), 'Artificial transmutation of the gene', *Science* **66**(1699), 84–87.

Nachman, M. W. (2001), 'Single nucleotide polymorphisms and recombination rate in humans', *TRENDS in Genetics* **17**(9), 481–485.

Nei, M. (1973), 'Analysis of gene diversity in subdivided populations', *Proceedings of the National Academy of Sciences* **70**(12), 3321–3323.

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

Nei, M. & Gojobori, T. (1986), 'Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions', *Molecular biology and evolution* **3**(5), 418–426.

Nielsen, C. B., Cantor, M., Dubchak, I., Gordon, D. & Wang, T. (2010), 'Visualizing genomes: techniques and challenges', *Nature methods* **7**(3s), S5.

Nielsen, R. (2005), 'Molecular signatures of natural selection', *Annu. Rev. Genet.* **39**, 197–218.

Ohta, T. (1973), 'Slightly deleterious mutant substitutions in evolution', *Nature* **246**(5428), 96–98.

Ohta, T. & Gillespie, J. H. (1996), 'Development of neutral and nearly neutral theories', *Theoretical population biology* **49**(2), 128–142.

Ortiz-Barrientos, D., Chang, A. S. & Noor, M. A. (2006), 'A recombinational portrait of the drosophila pseudoobscura genome', *Genetics Research* **87**(1), 23–31.

Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M. & Andolfatto, P. (2010), 'On the utility of short intron sequences as a reference for the detection of positive and negative selection in drosophila', *Molecular biology and evolution* **27**(6), 1226–1234.

Pearson, K. (1895), 'Note on regression and inheritance in the case of two parents', *Proceedings of the Royal Society of London* **58**, 240–242.

Pearson, W. R. & Lipman, D. J. (1988), 'Improved tools for biological sequence comparison', *Proceedings of the National Academy of Sciences* **85**(8), 2444–2448.

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. & Lercher, M. J. (2014), 'Popgenome: an efficient swiss army knife for population genomic analyses in r', *Molecular biology and evolution* **31**(7), 1929–1936.

Piganeau, G. & Eyre-Walker, A. (2003), 'Estimating the distribution of fitness effects from dna sequence data: implications for the molecular clock', *Proceedings of the National Academy of Sciences* **100**(18), 10335–10340.

Poliakov, A., Foong, J., Brudno, M. & Dubchak, I. (2014), 'Genomevista—an integrated software package for whole-genome alignment and visualization', *Bioinformatics* **30**(18), 2654–2655.

Pool, J. E. (2015), 'The mosaic ancestry of the drosophila genetic reference panel and the d. melanogaster reference genome reveals a network of epistatic fitness interactions', *Molecular biology and evolution* **32**(12), 3236–3251.

Pool, J. E., Corbett-Detig, R. B., Sugino, R. P., Stevens, K. A., Cardeno, C. M., Crepeau, M. W., Duchen, P., Emerson, J., Saelao, P., Begun, D. J. et al. (2012), 'Population genomics of sub-saharan drosophila melanogaster: African diversity and non-african admixture', *PLoS genetics* **8**(12), e1003080.

Pool, J. E. & Nielsen, R. (2008), 'The impact of founder events on chromosomal variability in multiply mating species', *Molecular biology and evolution* **25**(8), 1728–1736.

Powell, J. (1997), 'Progress and prospects in evolutionary biology. the drosophila model', *Oxford University Press, New York* .

Pybus, M., Dall'Olio, G. M., Luisi, P., Uzkudun, M., Carreno-Torres, A., Pavlidis, P., Laayouni, H., Bertranpetit, J. & Engelken, J. (2013), '1000 genomes selection browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans', *Nucleic acids research* **42**(D1), D903–D909.

Ràmia, M., Librado, P., Casillas, S., Rozas, J. & Barbadilla, A. (2011), 'Popdrowser: the population drosophila browser', *Bioinformatics* **28**(4), 595–596.

Ranz, J. M., Castillo-Davis, C. I., Meiklejohn, C. D. & Hartl, D. L. (2003), 'Sex-dependent gene expression and evolution of the drosophila transcriptome', *Science* **300**(5626), 1742–1745.

Rech, G. E., Bogaerts-Marquez, M., Barron, M. G., Merenciano, M., Villanueva-Canas, J. L., Horvath, V., Fiston-Lavier, A.-S., Luyten, I., Venkataram, S., Quesneville, H. et al. (2018), 'Stress response, behavior, and development are shaped by transposable element-induced mutations in drosophila', *bioRxiv* p. 380618.

Reeves, R. G., Bryk, J., Altrock, P. M., Denton, J. A. & Reed, F. A. (2014), 'First steps towards underdominant genetic transformation of insect populations', *PloS one* **9**(5), e97557.

Reinhardt, J. A., Kolaczkowski, B., Jones, C. D., Begun, D. J. & Kern, A. D. (2014), 'Parallel geographic variation in drosophila melanogaster', *Genetics* pp. genetics–114.

Revell, L. J. (2012), 'phytools: an r package for phylogenetic comparative biology (and other things)', *Methods in Ecology and Evolution* **3**(2), 217–223.

Rigden, D. J. & Fernández, X. M. (2017), 'The 2018 nucleic acids research database issue and the online molecular biology database collection', *Nucleic acids research* **46**(D1), D1–D7.

Rubin, G. M. (1996), 'Around the genomes: the drosophila genome project.', *Genome Research* **6**(2), 71–79.

Ràmia, M. (2015), Visualization, description and analysis of the genome variation of a natural population of Drosophila melanogaster, PhD thesis, Universitat Autònoma de Barcelona.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J. et al. (2002), 'Detecting recent positive selection in the human genome from haplotype structure', *Nature* **419**(6909), 832.

Sackton, T. B., Kulathinal, R. J., Bergman, C. M., Quinlan, A. R., Dopman, E. B., Carneiro, M., Marth, G. T., Hartl, D. L. & Clark, A. G. (2009), 'Population genomic inferences from sparse high-throughput sequencing of two populations of drosophila melanogaster', *Genome biology and evolution* **1**, 449–465.

Saitou, N. & Nei, M. (1987), 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Molecular Biology and Evolution* **4**(4), 406–425.

Salvador-Martínez, I., Coronado-Zamora, M., Castellano, D., Barbadilla, A. & Salazar-Ciudad, I. (2017), 'Mapping selection within drosophila melanogaster embryo's anatomy', *Molecular biology and evolution* **35**(1), 66–79.

Sanger, F. & Coulson, A. R. (1975), 'A rapid method for determining sequences in dna by primed synthesis with dna polymerase', *Journal of molecular biology* **94**(3), 441–448.

Schattner, P. (2008), *Genomes, Browsers and Databases. Data-Mining Tools for Integrated Genomic Databases*, Cambridge University Press, New York.

Schlotterer, C., Neumeier, H., Sousa, C. & Nolte, V. (2005), 'Highly structured asian d. melanogaster populations provide an effective tool for hitchhiking mapping', *Genetics* .

Sella, G., Petrov, D. A., Przeworski, M. & Andolfatto, P. (2009), 'Pervasive natural selection in the drosophila genome?', *PLoS genetics* **5**(6), e1000495.

Shapiro, J. A., Huang, W., Zhang, C., Hubisz, M. J., Lu, J., Turissini, D. A., Fang, S., Wang, H.-Y., Hudson, R. R., Nielsen, R. et al. (2007), 'Adaptive genic evolution in the drosophila genomes', *Proceedings of the National Academy of Sciences* **104**(7), 2271–2276.

Signor, S. A., New, F. N. & Nuzhdin, S. (2018), 'A large panel of drosophila simulans reveals an abundance of common variants', *Genome Biology and Evolution* **10**(1), 189–206.

Singh, N. D., Jensen, J. D., Clark, A. G. & Aquadro, C. F. (2012), 'Inferences of demography and selection in an african population of d. melanogaster', *Genetics* pp. genetics–112.

Singh, R. S. & Rhomberg, L. R. (1987), 'A comprehensive study of genic variation in natural populations of drosophila melanogaster. ii. estimates of heterozygosity and patterns of geographic differentiation', *Genetics* **117**(2), 255–271.

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J. & Holmes, I. H. (2009), 'Jbrowse: A next-generation genome browser', *Genome Research* **19**(9), 1630–1638.

Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E., Bult, C. J. & Group, M. G. D. (2017), 'Mouse genome database (mgd)-2018: knowledgebase for the laboratory mouse', *Nucleic acids research* **46**(D1), D836–D842.

Smith, J. M. & Haigh, J. (1974), 'The hitch-hiking effect of a favourable gene', *Genetics Research* **23**(1), 23–35.

Smith, N. G. & Eyre-Walker, A. (2002), 'Adaptive protein evolution in drosophila', *Nature* **415**(6875), 1022.

Spearman, C. (1904), 'The proof and measurement of association between two things', *The American journal of psychology* **15**(1), 72–101.

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A. et al. (2002), 'The generic genome browser: a building block for a model organism system database', *Genome research* **12**(10), 1599–1610.

Stephan, W. & Li, H. (2007), 'The recent demographic and adaptive history of drosophila melanogaster', *Heredity* **98**(2), 65.

Stoletzki, N. & Eyre-Walker, A. (2011), 'Estimation of the neutrality index', *Molecular biology and evolution* **28**(1), 63–70.

Tajima, F. (1989), 'Statistical method for testing the neutral mutation hypothesis by dna polymorphism.', *Genetics* **123**(3), 585–595.

Tajima, F. (1993), 'Measurement of dna polymorphism', *Mechanisms of molecular evolution* pp. 37–59.

Tajima, F. (1996), 'The amount of dna polymorphism maintained in a finite population when the neutral mutation rate varies among sites', *Genetics* **143**(3), 1457–1465.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013), 'Mega6: Molecular evolutionary genetics analysis version 6.0', *Molecular Biology and Evolution* **30**(12), 2725–2729.

Team, R. C. et al. (2013), 'R: A language and environment for statistical computing'.

Telonis-Scott, M. & Hoffmann, A. A. (2018), 'Enhancing ebony? common associations with a cis-regulatory haplotype for drosophila melanogaster thoracic pigmentation in a japanese population and australian populations', *Frontiers in physiology* **9**.

Theußl, S. & Zeileis, A. (2009), 'Collaborative software development using r-forge. special invited paper on" the future of r"', *The R Journal* **1**(1), 9–14.

Thompson, E. (1975), 'The estimation of pairwise relationships', *Annals of human genetics* **39**(2), 173–188.

Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. (2013), 'Integrative genomics viewer (igv): high-performance genomics data visualization and exploration', *Briefings in bioinformatics* **14**(2), 178–192.

Tora, L. (2002), 'A unified nomenclature for tata box binding protein (tbp)-associated factors (tafs) involved in rna polymerase ii transcription', *Genes & development* **16**(6), 673–675.

Tyler-Smith, C., Yang, H., Landweber, L. F., Dunham, I., Knoppers, B. M., Donnelly, P., Mardis, E. R., Snyder, M. & McVean, G. (2015), 'Where next for genetics and genomics?', *PLoS biology* **13**(7), e1002216.

Vicoso, B. & Charlesworth, B. (2006), 'Evolution on the x chromosome: unusual patterns and processes', *Nature Reviews Genetics* **7**(8), 645.

Vitti, J. J., Grossman, S. R. & Sabeti, P. C. (2013), 'Detecting natural selection in genomic data', *Annual review of genetics* **47**, 97–120.

Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. (2006), 'A map of recent positive selection in the human genome', *PLoS biology* **4**(3), e72.

Wall, J. D., Andolfatto, P. & Przeworski, M. (2002), 'Testing models of selection and demography in drosophila simulans', *Genetics* **162**(1), 203–216.

Wang, J., Kong, L., Gao, G. & Luo, J. (2013), 'A brief introduction to web-based genome browsers', *Briefings in Bioinformatics* **14**(2), 131–143.

Watterson, G. (1975), 'On the number of segregating sites in genetical models without recombination', *Theoretical Population Biology* **7**(2), 256 – 276.

Wickham, H. (2016), *ggplot2: elegant graphics for data analysis*, Springer.

Wilcoxon, F. (1945), 'Individual comparisons by ranking methods', *Biometrics bulletin* **1**(6), 80–83.

Wright, S. (1931), 'Evolution in mendelian populations', *Genetics* **16**(2), 97–159.

Xie, Y. (2012), 'knitr: elegant, flexible and fast dynamic report generation with r'.

Yang, Z. & Bielawski, J. P. (2000), 'Statistical methods for detecting molecular adaptation', *Trends in ecology & evolution* **15**(12), 496–503.

Zuckerkandl, E. & Pauling, L. (1965), Evolutionary divergence and convergence in proteins, *in* 'Evolving genes and proteins', Elsevier, pp. 97–166.

7. Annexes

7. Annexes

7.1 Supplementary Tables

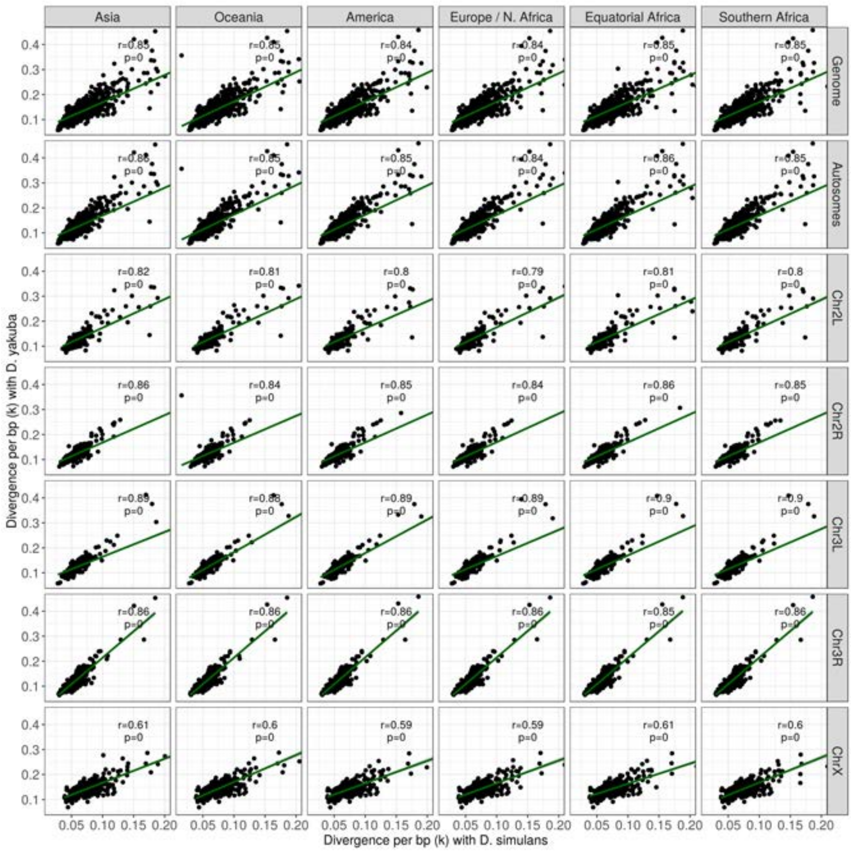
Supplementary Table 7.1: Adaptation metrics for genes grouped by recombination rate (ρ). Recombination estimates correspond to each population. Results obtained using the iMKT method.

	Bin	**Adaptation** ($\alpha_{asymptotic}$)	d**: strongly deleterious**	f**: neutral**	b**: weakly deleterious**	**mean** ρ **(± SD)**
Asia	1	0.3583	0.80632	0.19368	-	0.00016 (± 0.00011)
	2	0.4735	0.84429	0.08099	0.07472	0.00068 (± 0.00018)
	3	0.5385	0.84862	0.09071	0.06067	0.00147 (± 0.00027)
	4	0.5902	0.87828	0.07536	0.04636	0.0025 (± 3e-04)
	5	0.6103	0.87105	0.07185	0.0571	0.00383 (± 0.00049)
	6	0.6249	0.88222	0.06935	0.04843	0.00539 (± 0.00046)
	7	0.6516	0.88766	0.06004	0.0523	0.00743 (± 8e-04)
	8	0.5986	0.87561	0.07208	0.05231	0.01019 (± 0.00105)
	9	0.631	0.89015	0.07076	0.03909	0.01451 (± 0.00165)
	10	0.6642	0.89935	0.05478	0.04587	0.03034 (± 0.01603)
Oceania	1	0.3636	0.77654	0.10398	0.11947	0.00047 (± 0.00033)
	2	0.5509	0.82402	0.06351	0.11247	0.00167 (± 0.00037)
	3	0.6289	0.87262	0.05826	0.06911	0.00328 (± 0.00059)
	4	0.6133	0.87763	0.06174	0.06063	0.00497 (± 0.00042)
	5	0.625	0.88677	0.05758	0.05565	0.00676 (± 0.00057)
	6	0.7022	0.88165	0.05711	0.06124	0.0086 (± 0.00054)
	7	0.7157	0.88784	0.05407	0.05809	0.01067 (± 0.00054)
	8	0.6766	0.90004	0.05272	0.04724	0.01288 (± 7e-04)
	9	0.6621	0.88871	0.05914	0.05215	0.01636 (± 0.00126)
	10	0.678	0.8898	0.06013	0.05007	0.0258 (± 0.00662)
America	1	0.3198	0.74514	0.08737	0.16749	0.00122 (± 0.00086)
	2	0.4549	0.80048	0.06856	0.13095	0.00427 (± 9e-04)
	3	0.5893	0.84824	0.06163	0.09013	0.00732 (± 0.00082)
	4	0.591	0.85049	0.0647	0.08481	0.00997 (± 0.00062)
	5	0.6178	0.85819	0.06014	0.08167	0.01219 (± 0.00068)
	6	0.5985	0.86857	0.0576	0.07383	0.01463 (± 0.00072)
	7	0.6521	0.87632	0.05611	0.06757	0.0171 (± 0.00075)
	8	0.7151	0.88006	0.05437	0.06557	0.02021 (± 0.00102)
	9	0.6976	0.88497	0.05139	0.06364	0.0243 (± 0.00133)
	10	0.6677	0.88173	0.05662	0.06165	0.034 (± 0.00595)
Europe / North Africa	1	0.2706	0.73691	0.08541	0.17768	0.00164 (± 0.00121)
	2	0.3826	0.80181	0.07097	0.12723	0.00583 (± 0.00103)
	3	0.5675	0.84556	0.06035	0.09409	0.00881 (± 8e-04)
	4	0.6001	0.85417	0.06011	0.08572	0.01087 (± 0.00055)
	5	0.6325	0.8655	0.05437	0.08013	0.01289 (± 0.00057)
	6	0.6468	0.86868	0.05462	0.0767	0.01526 (± 0.00063)
	7	0.6749	0.87433	0.05301	0.07266	0.01781 (± 0.00078)
	8	0.6757	0.87764	0.0531	0.06926	0.02066 (± 0.00079)
	9	0.6786	0.88279	0.05013	0.06707	0.02492 (± 0.00159)
	10	0.7201	0.86751	0.05871	0.07378	0.03514 (± 0.00732)
Equatorial Africa	1	0.2602	0.71296	0.04686	0.24018	0.00536 (± 0.00319)
	2	0.4662	0.77647	0.04883	0.17469	0.0161 (± 0.00311)
	3	0.5438	0.81522	0.04673	0.13805	0.02473 (± 0.00222)
	4	0.6165	0.83702	0.0401	0.12288	0.03222 (± 0.00226)
	5	0.6625	0.84168	0.04057	0.11775	0.04179 (± 0.00301)
	6	0.6606	0.85083	0.04125	0.10791	0.05177 (± 0.00294)
	7	0.702	0.86782	0.03479	0.09739	0.0635 (± 0.00395)
	8	0.6967	0.85226	0.04273	0.10501	0.07852 (± 0.0047)
	9	0.7589	0.8681	0.03657	0.09532	0.10151 (± 0.00696)
	10	0.7754	0.87413	0.03522	0.09065	0.14461 (± 0.03086)
Southern Africa	1	0.1423	0.75401	0.07764	0.16836	0.00356 (± 0.00249)
	2	0.5422	0.83507	0.04375	0.12118	0.01313 (± 0.00253)
	3	0.6189	0.85935	0.04355	0.09711	0.02164 (± 0.00202)
	4	0.6283	0.86373	0.04135	0.09492	0.02801 (± 0.00194)
	5	0.6684	0.87652	0.04007	0.08341	0.03554 (± 0.00211)
	6	0.6667	0.87946	0.04072	0.07982	0.04287 (± 0.0022)
	7	0.6559	0.8788	0.04163	0.07957	0.05229 (± 0.00335)
	8	0.6676	0.88482	0.03905	0.07613	0.06529 (± 0.00479)
	9	0.751	0.88319	0.0395	0.07732	0.10018 (± 0.02055)
	10	0.8212	0.88983	0.03584	0.07433	0.22418 (± 0.06106)

Supplementary Table 7.2: Adaptation metrics for genes grouped according to coding density. Coding density metrics estimated based on the reference genome. Results obtained using the iMKT method.

	Bin	Adaptation ($\alpha_{asymptotic}$)	d: strongly deleterious	f: neutral	b: weakly deleterious	mean coding density (± SD)
Asia	1	0.6697	0.86667	0.07251	0.06083	0.0677 (±0.02585)
	2	0.5745	0.8651	0.08312	0.05178	0.12719 (±0.01008)
	3	0.5929	0.87497	0.0756	0.04943	0.16111 (±0.01028)
	4	0.6123	0.87428	0.07443	0.05128	0.19689 (±0.00912)
	5	0.6096	0.87571	0.0674	0.05688	0.22361 (±0.00857)
	6	0.6779	0.88341	0.06457	0.05202	0.25689 (±0.0105)
	7	0.6004	0.89055	0.06809	0.04136	0.29408 (±0.01128)
	8	0.5983	0.87395	0.07454	0.05151	0.32773 (±0.01011)
	9	0.5268	0.87385	0.0757	0.05045	0.36959 (±0.01554)
	10	0.6373	0.88865	0.05693	0.05442	0.46205 (±0.04944)
Oceania	1	0.6712	0.87089	0.06764	0.06147	0.0677 (±0.02585)
	2	0.6288	0.87294	0.06877	0.0583	0.12719 (±0.01008)
	3	0.6733	0.87859	0.05405	0.06737	0.16111 (±0.01028)
	4	0.6389	0.87927	0.0632	0.05752	0.19689 (±0.00912)
	5	0.6515	0.87905	0.05773	0.06322	0.22361 (±0.00857)
	6	0.6879	0.88754	0.05474	0.05772	0.25689(±0.0105)
	7	0.6635	0.89308	0.05131	0.05561	0.29408 (±0.01128)
	8	0.6665	0.8782	0.05204	0.06976	0.32773 (±0.01011)
	9	0.6034	0.87698	0.05675	0.06627	0.36959 (±0.01554)
	10	0.6411	0.88735	0.05669	0.05596	0.46205 (±0.04944)
America	1	0.6488	0.84961	0.06713	0.08325	0.0677 (±0.02585)
	2	0.6211	0.85845	0.0658	0.07574	0.12719 (±0.01008)
	3	0.6356	0.8646	0.05986	0.07555	0.16111 (±0.01028)
	4	0.6295	0.86212	0.06044	0.07745	0.19689 (±0.00912)
	5	0.6346	0.86119	0.05822	0.0806	0.22361 (±0.00857)
	6	0.6611	0.86959	0.05645	0.07395	0.25689(±0.0105)
	7	0.6619	0.87822	0.04961	0.07216	0.29408 (±0.01128)
	8	0.6092	0.85695	0.06081	0.08225	0.32773 (±0.01011)
	9	0.5575	0.86399	0.05901	0.07701	0.36959 (±0.01554)
	10	0.634	0.87304	0.0549	0.07206	0.46205 (±0.04944)
Europe / N. Africa	1	0.6687	0.84789	0.0626	0.08951	0.0677 (±0.02585)
	2	0.6195	0.85217	0.0636	0.08423	0.12719 (±0.01008)
	3	0.6383	0.86002	0.05846	0.08152	0.16111 (±0.01028)
	4	0.6333	0.86214	0.05893	0.07893	0.19689 (±0.00912)
	5	0.6305	0.8572	0.05804	0.08476	0.22361 (±0.00857)
	6	0.6639	0.86919	0.05343	0.07738	0.25689(±0.0105)
	7	0.6629	0.87378	0.04826	0.07797	0.29408 (±0.01128)
	8	0.622	0.8551	0.05527	0.08963	0.32773 (±0.01011)
	9	0.564	0.86022	0.05689	0.08289	0.36959 (±0.01554)
	10	0.6269	0.87283	0.05371	0.07346	0.46205 (±0.04944)
Equatorial Africa	1	0.6823	0.8311	0.04557	0.12333	0.0677 (±0.02585)
	2	0.647	0.8378	0.04494	0.11726	0.12719 (±0.01008)
	3	0.6773	0.84446	0.04089	0.11465	0.16111 (±0.01028)
	4	0.6684	0.84724	0.04241	0.11035	0.19689 (±0.00912)
	5	0.6757	0.84579	0.03857	0.11563	0.22361 (±0.00857)
	6	0.7055	0.85442	0.03612	0.10946	0.25689(±0.0105)
	7	0.6856	0.85782	0.03499	0.1072	0.29408 (±0.01128)
	8	0.635	0.84383	0.04041	0.11577	0.32773 (±0.01011)
	9	0.6109	0.84393	0.03995	0.11612	0.36959 (±0.01554)
	10	0.6625	0.85689	0.03683	0.10628	0.46205 (±0.04944)
Southern Africa	1	0.6833	0.86123	0.04787	0.09091	0.0677 (±0.02585)
	2	0.663	0.86444	0.04475	0.0908	0.12719 (±0.01008)
	3	0.6795	0.87289	0.04211	0.085	0.16111 (±0.01028)
	4	0.672	0.87114	0.04375	0.08511	0.19689 (±0.00912)
	5	0.701	0.87745	0.03765	0.0849	0.22361 (±0.00857)
	6	0.716	0.8813	0.03761	0.08109	0.25689(±0.0105)
	7	0.7007	0.88369	0.03494	0.08136	0.29408 (±0.01128)
	8	0.6512	0.86977	0.04086	0.08937	0.32773 (±0.01011)
	9	0.6188	0.87158	0.04158	0.08684	0.36959 (±0.01554)
	10	0.6675	0.87953	0.03895	0.08152	0.46205 (±0.04944)

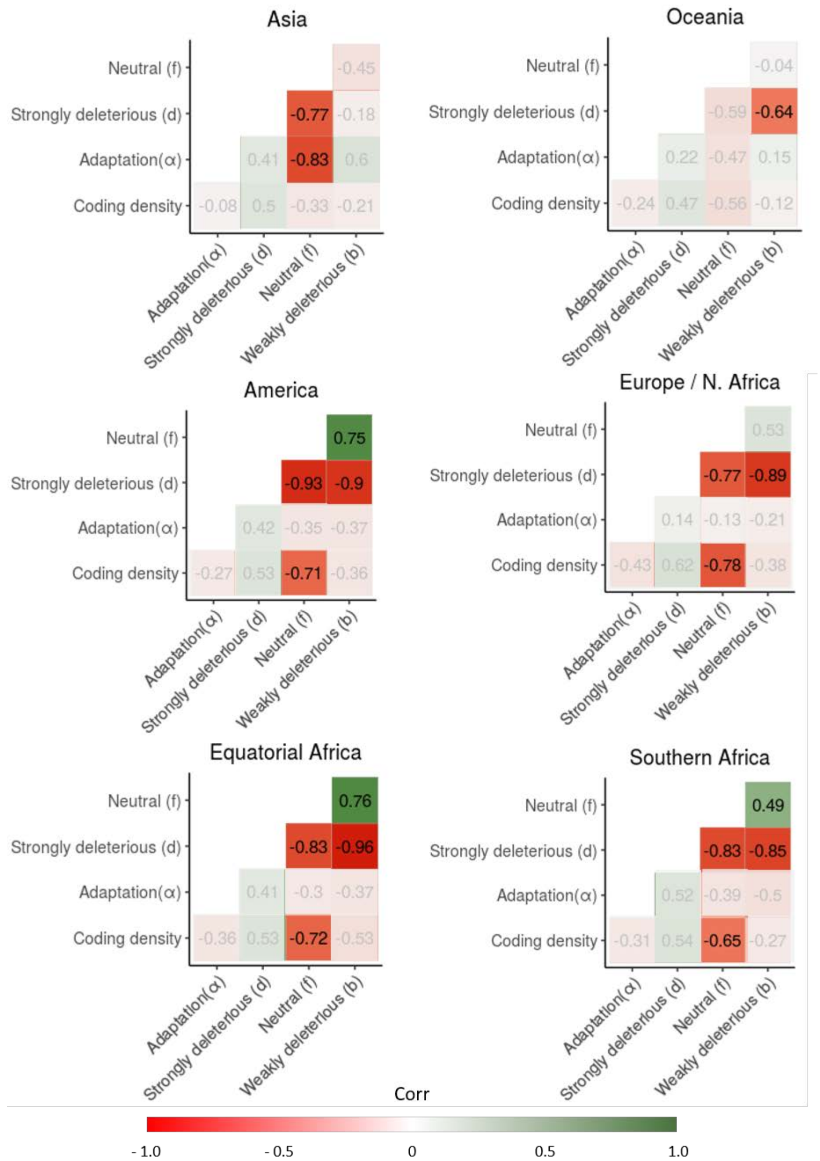
7.2 Supplementary Figures



Supplementary Figure 7.1: Correlation between divergence metrics using *D. simulans* and *D. yakuba*. Spearman's correlation coefficients (Spearman 1904), with their associated p-values. Note that values in the Y axis are two times higher than in the X axis. All situations show a positive correlation between both parameters, with an associated $p = 0$.

Supplementary Figure 7.2: Correlation between polymorphism and (A) historical recombination and (B) coding density. Spearman's correlation coefficient with its associated $p-value$. In general, we observe a positive and significant correlation between π and recombination, but this correlation is not present between π and coding density.

Supplementary Figure 7.3: Correlation between divergence and (A) historical recombination and (B) coding density. Spearman's correlation coefficient with its associated $p - value$. In general, we do not observe any significant correlation between divergence and neither recombination nor coding density.



Supplementary Figure 7.4: Spearman's correlation coefficients between adaptation metrics and coding density. Coding density, rate of adaptive evolution (α) and fraction of strongly deleterious (d), neutral (f) and weakly deleterious sites (b) for 13,754 genes grouped in ten bins based on coding density estimates. Correlation coefficients with an associated $p > 0.05$ are shaded in grey.

7.3 Article 1. PopFly: the *Drosophila* population genomics browser

Sergi Hervas, Esteve Sanz, Sònia Casillas, John E. Pool, and Antonio Barbadilla (2017) PopFly: the *Drosophila* population genomics browser, Bioinformatics 33(17): 2779–2780. `https://doi.org/10.1093/bioinformatics/btx301`. PubMed Central PMCID: PMC5860067.

Bioinformatics, 33(17), 2017, 2779–2780
doi: 10.1093/bioinformatics/btx301
Advance Access Publication Date: 4 May 2017
Applications Note

OXFORD

Genetics and population analysis

PopFly: the *Drosophila* population genomics browser

Sergi Hervas[1,*], Esteve Sanz[2], Sònia Casillas[1], John E. Pool[3] and Antonio Barbadilla[1,*]

[1]Institut de Biotecnologia i de Biomedicina and Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain, [2]Servei de Genòmica i Bioinformàtica, Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain and [3]Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on February 24, 2017; revised on April 27, 2017; editorial decision on May 1, 2017; accepted on May 2, 2017

Abstract

Summary: The recent compilation of over 1100 worldwide wild-derived *Drosophila melanogaster* genome sequences reassembled using a standardized pipeline provides a unique resource for population genomic studies (Drosophila Genome Nexus, DGN). A visual display of the estimated metrics describing genome-wide variation and selection patterns would allow gaining a global view and understanding of the evolutionary forces shaping genome variation.

Availability and implementation: Here, we present PopFly, a population genomics-oriented genome browser, based on JBrowse software, that contains a complete inventory of population genomic parameters estimated from DGN data. This browser is designed for the automatic analysis and display of genetic variation data within and between populations along the *D. melanogaster* genome. PopFly allows the visualization and retrieval of functional annotations, estimates of nucleotide diversity metrics, linkage disequilibrium statistics, recombination rates, a battery of neutrality tests, and population differentiation parameters at different window sizes through the euchromatic chromosomes. PopFly is open and freely available at site http://popfly.uab.cat.

Contact: sergi.hervas@uab.cat or antonio.barbadilla@uab.cat

1 Introduction

High-throughput sequencing technologies are allowing the description of genome-wide variation patterns of an ever growing number of organisms. Several studies have been carried out in the last years involving dozens and even hundreds of wild-derived samples in the species *Drosophila melanogaster*, the model organism for population genetic studies (reviewed by Casillas and Barbadilla, 2017). The Drosophila Genome Nexus (DGN) project has reassembled most published *D. melanogaster* population genomic data, creating a set of around 1100 worldwide genome sequences comparable among them, which greatly facilitates future population genomic studies in this model species (Lack *et al.*, 2015, 2016).

One main bioinformatics challenge when analyzing a huge amount of genomic data is how to get an easy and intuitive visualization and retrieval of such information. Genome browsers provide a unique platform for molecular biologists to browse, search, retrieve and analyze these genomic data efficiently and conveniently taking advantage of their user-friendly graphical interface. PopDrowser, the Population *Drosophila* Browser (Ràmia *et al.*, 2012), displays population genomic parameters estimated from a single population of *D. melanogaster* (the Drosophila Genetic Reference Panel, Mackay *et al.*, 2012). However, this browser has become outdated in terms of performance and data storage. Here, we present PopFly, a population genomics-oriented web-browser that updates our previous PopDrowser. PopFly contains a complete inventory of population genomic parameters estimated from the DGN project data, along with functional annotations from the reference *D. melanogaster* genome sequence. The user-friendly graphical web interface of this new browser allows an easy visualization and retrieval of the broadest catalog of genome-wide patterns of nucleotide variation and population genetics estimates in *D. melanogaster* at different resolution scales. Furthermore, the automated nature of the data processing pipeline makes this platform highly scalable, allowing the continuous updating of the database by the addition of the increasing number of new genome sequences.

2 Browser overview

2.1 Input data

The input data is a set of aligned *D. melanogaster* genome sequences from the DGN project. At present, the analyzed data comprises more than 960 genome sequences from 30 populations out of 18 countries spanning 5 continents. The genome sequences of *Drosophila yakuba* and *Drosophila simulans* are used as outgroup species.

2.2 Software implementation

PopFly contains a set of precomputed population genomic estimates generated through the combined implementation of programs VariScan2 (Hutter *et al.*, 2006), LDhelmet (Chan *et al.*, 2012), and custom ad-hoc scripts. Data and summary statistics are graphically displayed along the chromosome arms on a web-based user interface (Fig. 1) using the JBrowse software (Buels *et al.*, 2016), which considerably improves the performance of its previous version, and can be easily downloaded in bedGraph, wiggle or gff3 format files. PopFly also incorporates utilities to perform on-the-fly statistical analyses and download sequences, and allows uploading user custom tracks. The current browser implementation is running under Apache on a CentOS 7.2 Linux x64 server, 16 IntelXeon 2.4GHz processors, 32GB RAM.

2.3 Browser tracks

The genome browser includes, for each sampled population and metapopulation (populations aggregated by continent): summary measures of nucleotide diversity, divergence between species, linkage disequilibrium statistics, historical population-scaled recombination rate estimates, a battery of neutrality tests and population differentiation metrics (Table 1), computed at non-overlapping windows of varying size (1 kb, 10 kb, 50 kb, 100 kb). The browser also contains the *D. melanogaster* genome reference sequence along with its functional annotations (version 5.57 from FlyBase), and the high-resolution reference recombination maps from Comeron *et al.*, (2012) and Fiston-Lavier *et al.*, (2010).

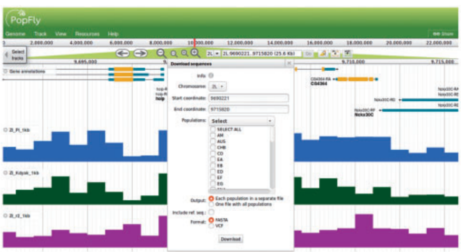


Fig. 1. PopFly snapshot with certain activated tracks and the utility to download sequences

Acknowledgements

We would like to thank Miquel Ràmia for his helpful guidance and suggestions. We also thank Josefa González and the JBrowse developing community for their valuable comments to improve PopFly.

Funding

This work was supported by the Ministerio de Economía y Competitividad [BFU2013-42649-P to A.B.]; the Generalitat de Catalunya [2014-SGR-1346]; the Departament de Genètica i Microbiologia of the Universitat Autònoma de Barcelona [12ª PIPF to S.H.]; the Youth Employment Initiative and European Social Fund [PEJ-2014 to E.S]; and the National Institutes of Health [R01 GM111797 to J.E.P.].

Conflict of Interest: none declared.

References

Buels,R. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.

Casillas,S. and Barbadilla,A. (2017) Molecular population genetics. *Genetics*, **205**, 1003–1035.

Chan,A.H. *et al.* (2012) Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. *PLoS Genet.*, **8**, e1003090.

Comeron,J.M. *et al.* (2012) The many landscapes of recombination in Drosophila melanogaster. *PLoS Genet.*, **8**, e1002905.

Fiston-Lavier,A.-S. *et al.* (2010) Drosophila melanogaster recombination rate calculator. *Gene*, **463**, 18–20.

Hutter,S. *et al.* (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinform.*, **7**, 409.

Lack,J.B. *et al.* (2016) A thousand fly genomes: an expanded drosophila genome nexus. *Mol. Biol. Evol.*, **33**, 3308–3313.

Lack,J.B. *et al.* (2015) The drosophila genome nexus: a population genomic resource of 623 drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics*, **199**, 1229–1241.

Mackay,T.F.C. *et al.* (2012) The drosophila melanogaster genetic reference panel. *Nature*, **482**, 173–178.

Ràmia,M. *et al.* (2012) PopDrowser: the population drosophila browser. *Bioinformatics*, **28**, 595–596.

Table 1. PopFly category tracks

Category	Annotations and main parameter estimates
Reference tracks	*D. melanogaster* reference genome (build 5.57) sequence and annotations
Frequency-based nucleotide variation	Watterson's nucleotide diversity (θ), nucleotide diversity (π), number of 0-fold and 4-fold segregating sites (P_{0f}, P_{4f}), 0-fold and 4-fold nucleotide diversity (π_{0f}, π_{4f})
Divergence-based metrics	Nucleotide divergence per bp (k) with *D. yakuba* and *D. simulans*, number of 0-fold and 4-fold divergent sites (D_{0f}, D_{4f}), and 0-fold and 4-fold divergence (k_{0f}, k_{4f})
Linkage disequilibrium	LD sites, D, |D|, D', |D'|, r^2, number of haplotypes (h), haplotype diversity (Hd)
Recombination	Recombination rate estimates from Comeron *et al.* (2012) and Fiston-Lavier *et al.* (2010), historical population-scaled recombination rate (ρ_A=2N_er; ρ_X = 8/3 N_er)
Selection tests based on SFS and/or variability	Fu & Li D and F test statistics, Tajima's D, Fu's Fs statistic
Selection tests based on polymorphism and divergence	Ka/Ks ratio, neutrality index (NI), direction of selection (DoS), proportion of adaptive substitutions (α) from McDonald-Kreitman test
Population differentiation	F_{ST} estimates between populations

7.4 Article 2. PopHuman: the human population genomics browser

Sònia Casillas, Roger Mulet, Pablo Villegas-Mirón, Sergi Hervas, Esteve Sanz, Daniel Velasco, Jaume Bertranpetit, Hafid Laayouni, and Antonio Barbadilla (2018) PopHuman: the human population genomics browser. Nucleic Acids Research 46(D1):D1003-D1010. https://doi.org/10.1093/nar/gkx943. PubMed PMID: 29059408; PubMed Central PMCID: PMC5753332.

Published online 20 October 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D1003–D1010***
doi: 10.1093/nar/gkx943

PopHuman: the human population genomics browser

Sònia Casillas[1,*,†], Roger Mulet[1,†], Pablo Villegas-Mirón[2], Sergi Hervas[1], Esteve Sanz[3], Daniel Velasco[1], Jaume Bertranpetit[2], Hafid Laayouni[2,4] and Antonio Barbadilla[1,3,*]

[1]Institut de Biotecnologia i de Biomedicina and Department de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain, [2]Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Doctor Aiguader 88 (PRBB), 08003 Barcelona, Catalonia, Spain, [3]Servei de Genòmica i Bioinformàtica, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain and [4]Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, 08003 Barcelona, Spain

Received August 11, 2017; Revised September 18, 2017; Editorial Decision October 02, 2017; Accepted October 04, 2017

ABSTRACT

The 1000 Genomes Project (1000GP) represents the most comprehensive world-wide nucleotide variation data set so far in humans, providing the sequencing and analysis of 2504 genomes from 26 populations and reporting >84 million variants. The availability of this sequence data provides the human lineage with an invaluable resource for population genomics studies, allowing the testing of molecular population genetics hypotheses and eventually the understanding of the evolutionary dynamics of genetic variation in human populations. Here we present PopHuman, a new population genomics-oriented genome browser based on JBrowse that allows the interactive visualization and retrieval of an extensive inventory of population genetics metrics. Efficient and reliable parameter estimates have been computed using a novel pipeline that faces the unique features and limitations of the 1000GP data, and include a battery of nucleotide variation measures, divergence and linkage disequilibrium parameters, as well as different tests of neutrality, estimated in non-overlapping windows along the chromosomes and in annotated genes for all 26 populations of the 1000GP. PopHuman is open and freely available at http://pophuman.uab.cat.

INTRODUCTION

Soon after the elucidation of the entire human genome (1–3), the description of genetic variation in human populations and the identification of those variants that affect health and disease became the next challenges of genomics research (4). The International HapMap Consortium built the first genome-wide catalog of common human genetic variation in diverse populations (4–6), charting haplotype maps of 1.6 million single nucleotide polymorphisms (SNPs) in 1184 reference individuals from 11 global populations. In addition to numerous genome-wide association studies (GWAS) (7), the HapMap data allowed the detection of positive natural selection across the human genome (8,9), as well as the development of new tests to infer recent episodes of selective sweeps based on the length of haplotypes, such as the Long-Range Haplotype (LRH) (10), the integrated Haplotype Score (iHS) (11), and the Cross Population Extended Haplotype Homozygosity (XP-EHH) (8).

During the last decade, the development of next generation sequencing (NGS) technologies (12,13) has allowed the deciphering of complete genome sequences of thousands of human individuals, and the 1000 Genomes Project (1000GP) has become the reference data set for population genetics and genomics (14,15). With the aim of providing a deep characterization of human genome sequence variation, the most recent version of the 1000GP (Phase III) completes the sequencing and analysis of 2504 genomes from 26 populations and describes most variants with frequencies as low as 1%. Due to its higher resolution and smaller SNP ascertainment bias compared to HapMap genotyping data, the availability of the 1000GP data provides the human lineage with an invaluable resource on which to test molecular population genetics hypotheses and eventually understand the evolutionary dynamics of genetic variation in human populations (16).

Regions of the genome that are (or have been) subject to natural selection show distinctive patterns of genetic variation in the DNA sequence (17). The signature of long-range haplotypes persists for a relatively short period of time (<30 000 years), and related statistics can detect very recent selection only. However, other signatures persist longer in the genome: differentiation between populations (<50 000–<75 000 years), high frequency derived alleles (<80 000 years), reduction in genetic diversity and excess of rare al-

*To whom correspondence should be addressed. Sònia Casillas. Tel: +34 93 5868958; Fax: +34 93 5812011; Email: sonia.casillas@uab.cat
Correspondence may also be addressed to Antonio Barbadilla. Email: antonio.barbadilla@uab.cat
†These authors contributed equally to this work as first authors.
Present address: Roger Mulet, Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands.

D1004 Nucleic Acids Research, 2018, Vol. 46, Database issue

leles (<250 000 years), and high proportion of function-altering substitutions between species (many millions of years) (17).

Population genomics analyses of the 1000GP data set can be largely facilitated by (i) making an inventory of parameter values along the chromosomes that capture the evolutionary properties of the available sequences, and (ii) allowing the query and visualization of these estimates in a genome browser designed specifically for this data. As far as we are concerned, the 1000 Genomes Selection Browser 1.0 (18) is the only previous database that allows the interactive visualization and retrieval of population genetics metrics for the 1000GP data. It was published when the 1000GP was still in its first phase (1,092 individuals, 14 populations, 38 million SNPs) (14), and analyzed within-species polymorphism data for three populations in 30 kb sliding windows (18). Here, we present PopHuman, a new population genomics-oriented genome browser. PopHuman represents not only an update to the 1000GP Phase III (2504 individuals, 26 populations, 84.7 million SNPs), but also dramatic improvements in the amount of data analyzed and browser performance, compared to the 1000 Genomes Selection Browser 1.0. Furthermore, PopHuman analyzes between-species divergence, which allows the implementation of statistical tests to detect the signature of recurrent natural selection acting over prolonged periods of time, such as the McDonald and Kreitman test (MKT) (19), instead of recent selective sweeps only. Supplementary Table S1 details the differences between the two databases.

POPHUMAN ANALYSIS PIPELINE

We have designed and implemented a custom pipeline (Figure 1) facing the unique features and limitations of the 1000GP Phase III data (15). The pipeline discards reportedly inbred individuals (20) and non-accessible nucleotides (15), incorporates the genomic sequence of the chimpanzee (21) as outgroup, and estimates a battery of nucleotide variation, divergence and linkage disequilibrium parameters, as well as different tests of neutrality, on the filtered data. Several metrics have been computed both in non-overlapping sliding windows along the chromosomes and in annotated protein coding genes for 26 populations of distinct geographical origin (15).

Pre-processing of the 1000GP Phase III data

We retrieved human genome variation data generated by the 1000GP Phase III (15) from http://www.internationalgenome.org/data in Variant Call Format (VCF). This included 84.4 million variants detected across 2504 individuals from 26 different populations, mapped to the human reference genome version GRCh37/hg19. We want to warn the user that four of the analyzed populations present admixture (corresponding to the Admixed American metapopulation), so special care should be taken while interpreting PopHuman results in those cases.

Inbred individuals. The initial VCF files were filtered to exclude 243 individuals with inbreeding coefficients similar or greater than the ones expected for first-cousin offspring, according to Gazal *et al.* (20).

Genome accessibility mask. Due to the nature of short-read sequencing, sequencing depth varies along the length of the genome. The 1000GP provides an 'accessibility mask', a Browser Extensible Data (BED) file that indicates which sites of the genome were accessible to the sequencing techniques and have power for variant discovery (15). Two definitions were used in the Phase III, of which we selected the 'pilot-style' mask. This definition is less conservative than the 'strict' mask while being still adequate for population genomics analyses, and was chosen to maximize the amount of genomic sequence to be analyzed. It excludes the portion of the genome where depth of coverage (summed across all samples) was higher or lower than the average depth by a factor of 2-fold, as well as sites where >20% of overlapping reads had mapping quality of zero. Overall, 89.4% of the genome is considered reliable (95.9% of the non-N bases). Specifically, we placed 10 kb non-overlapping sliding windows in accessible regions of the genome (i.e. windows do not overlap any non-accessible nucleotide) to focus on high quality genomic regions only. Table 1 summarizes the total amount of data analyzed by PopHuman by following this methodology. In addition, we analyzed longer non-overlapping sliding windows of 100 kb placed all along the genome (i.e. windows might overlap non-accessible nucleotides, although these positions were discarded for the population genomics analyses) to focus on broader scale patterns of diversity across the genome.

Ancestral states. The ancestral states of human segregating sites were taken from the 1000GP Phase III (15), which were obtained by using the 6-way EPO alignments available in Ensembl v71 (22).

Outgroup species. To compute divergence metrics and neutrality tests based on the comparison of polymorphism and divergence, we added differences between humans and chimpanzees to the VCF files, as identified from a pre-computed hg19 => panTro4 alignment obtained from the VISTA browser (23) in multi-FASTA format (MFA). Specifically, the pairwise alignment was converted to VCF using custom scripts and merged with the 1000GP VCF files using *bcftools merge*.

Recombination

The most recent human genetic sex-specific maps were obtained from Bhérer *et al.* (24), based on a total of 104 246 informative meioses from six recent studies of human pedigrees.

Estimation of population genomics statistics

Windows-based. Several windows-based variation statistics and tests of neutrality (Table 2) were computed for each population separately using the R package PopGenome (25) and custom functions, considering biallelic SNPs as within-species variation data. Haplotype-based statistics (iHS and XP-EHH) were computed in a multithreaded

Nucleic Acids Research, 2018, Vol. 46, Database issue ***D1005***

Figure 1. PopHuman pipeline. Cited references in the figure: [1]1000GP Phase III (15); [2]Inbred individuals in the 1000GP (20); [3]VISTA Genome Browser (23); [4]Human genetic maps (24); [5]PopGenome software (25); [6]UCSC Genome Browser (35); [7]JBrowse software (34).

Table 1. Summary of the amount of data analyzed in PopHuman

Chromosome		Windows-based analysis		Genes-based analysis	
Chromosome number	Chromosome size (millions of bases)[a]	Number of windows[b]	Number of bases (millions)	Percentage of analyzed bases	Number of RefSeq[c] genes analyzed
1	249.25	14 741	147.41	59.14	2328
2	243.20	16 270	162.70	66.90	1464
3	198.02	13 575	135.75	68.55	1274
4	191.15	12 512	125.12	65.45	879
5	180.92	12073	120.73	66.73	1022
6	171.12	11 433	114.33	66.81	1206
7	159.14	9919	99.19	62.33	1108
8	146.36	9783	97.83	66.84	818
9	141.21	7358	73.58	52.11	944
10	135.53	8760	87.60	64.63	903
11	135.01	8877	88.77	65.75	1439
12	133.85	8773	87.73	65.54	1175
13	115.17	6481	64.81	56.27	449
14	107.35	5948	59.48	55.41	779
15	102.53	5334	53.34	52.02	791
16	90.35	4688	46.88	51.88	938
17	81.20	4556	45.56	56.11	1358
18	78.08	5164	51.64	66.14	341
19	59.13	2681	26.81	45.34	1609
20	63.03	4091	40.91	64.91	647
21	48.13	2211	22.11	45.94	296
22	51.30	2009	20.09	39.16	535
X	155.27	9312	93.12	59.97	918
Y	59.37	622	6.22	10.48	53
TOTAL	3095.68	187 171	1871.71	60.46	23 274

[a]Chromosome sizes are according to version GRCh37/hg19 of the human genome.
[b]Non-overlapping sliding windows of 10 kb have been defined such that they do not include non-accessible bases according to the Pilot-style Accessibility Mask of the 1000GP (15).
[c]RefSeq genes provided by the NCBI Entrez Gene database (33).

framework implemented by the program selscan (26), considering biallelic SNPs with Minor Allele Frequency (MAF) > 0.05 and a maximum gap of 20 kb between two consecutive SNPs. Then, whole chromosome per-SNP scores were summarized by calculating the mean of the absolute value of these scores for all SNPs in a window (27). Sexual chromosomes were not analyzed in these cases.

Genes-based. Comparisons of DNA polymorphism within populations and divergence to an outgroup species using the MKT (19) have been extensively used to detect the signature of natural selection at the molecular level (28). The MKT can be generalized to any two types of sites provided that one of them is assumed to evolve neutrally and that both types of sites are closely linked in the genome (29–31). Furthermore, Mackay *et al.* (32) developed an integrative new framework for the MKT by incorporating information on the MAF of the segregating sites, which allows estimating the fraction of new mutations that are strongly deleterious (and therefore not segregating), slightly deleterious (segregating at low frequency), old neutral (neutral before the split of humans and chimpanzees), and recently neutral (since the split of humans and chimpanzees), as well as the fraction of adaptive fixations. The standard and integrative MKTs (Table 3) were applied to all annotated human protein coding genes in RefSeq (33) and for different types of sites (i.e. 0-fold nonsynonymous coding sites, 5′UTR, 3′UTR, introns, and ±500 bp intergenic flanking regions, compared to 4-fold synonymous coding sites), for each population separately, using custom functions build within PopGenome (25).

OVERVIEW OF THE POPHUMAN GENOME BROWSER

PopHuman is a new population genomics-oriented genome browser based on JBrowse (34) that allows the interactive visualization and retrieval of several metrics estimated in non-overlapping sliding windows along the chromosomes and in annotated genes for all 26 populations of the 1000GP. It also includes a number of utilities and support resources.

JBrowse implementation

PopHuman is built on JBrowse (34) and is currently running under Apache on a CentOS 7.2 Linux x64 server with 16 Intel Xeon 2.4 GHz processors and 32 GB RAM.

Browser tracks

Variation statistics. Windows-based variation statistics and tests of neutrality (Table 2) are classified into: (i) frequency-based nucleotide variation; (ii) divergence-based metrics; (iii) linkage disequilibrium; (iv) recombination; (v) selection tests based on the Site Frequency Spectrum (SFS) and/or variability and (vi) selection tests based on the MKT. They are displayed for each population separately as histogram plots, with a yellow line showing the mean, and two shaded bands showing ±1 and ±2 standard deviations from the mean. Visualization style can be customized using the 'Edit config' option for each track.

Reference tracks. Several tracks have been imported from the UCSC Genome Browser (35) (Supplementary Table S2) and can be visualized along with variation statistics. They are classified into: (i) sequencing and annotation; (ii) regulation; (iii) comparative genomics; (iv) variation and (v) repeats.

Utilities and support resources

Tracks selector. PopHuman contains more than a thousand tracks, including both variation statistics (Table 2) and reference tracks (Supplementary Table S2). Given the large number of tracks available, these can be filtered and selected using the 'Select tracks' tool, which can be accessed from the top left corner, below the navigation bar. The filtering process is normally performed by first narrowing the search using the menu on the left, and then selecting the tracks of interest from the main panel on the right. This process can be done several times in order to finally get all the desired tracks selected.

Downloading raw data. Variation statistics for a given region can be conveniently downloaded in bedGraph, Wiggle or GFF3 formats using the 'Save track data' option for each track. In addition, bulk downloads of full variation tracks are available in BigWig format from the Resources menu. Finally, variant calls for the analyzed individuals can also be downloaded in VCF format using the PopHuman utility 'Download sequences', which can be accessed either from the Resources menu, or directly from the navigation bar.

Integrative MKT. Gene-based MKTs (Table 3) can be retrieved by right-clicking a gene and selecting the option 'Integrative MKT'.

Help section. The Help section contains exhaustive documentation about the 1000GP Phase III data analyzed by PopHuman and details about the browser tracks. Interestingly, it contains a comprehensive tutorial introducing to the usage of the database and to the testing of evolutionary hypotheses from a population genetics perspective. The tutorial works out, in different sequential steps, the visualization and analysis of a genomic region of around 20 kb in chromosome 7 that includes the *TRPV6* gene. *TRPV6* is a well-studied protein coding gene involved in the absorption of calcium from the diet that has experienced parallel selective sweeps in non-African populations, coinciding with the establishment of agriculture first in Europe around 10 000 years ago, and later in Asia. The tutorial contains several step-by-step guides to facilitate reproducing the results that are shown both in the form of figures and descriptive text.

Availability

All data, tools and support resources provided by PopHuman, as well as reference tracks downloaded from the UCSC Genome Browser (35), are open and freely available at http://pophuman.uab.cat.

COMPARISON TO OTHER DATABASES

While the PopHuman analysis pipeline presented here is completely novel, the genome browser is based on a similar

Table 2. List of major windows-based variation statistics and tests of neutrality in PopHuman, computed for each population separately

Category	Track name	Track description	Reference
Frequency-based nucleotide variation	S	Number of segregating sites per site	(42)
	Pi	Nucleotide diversity: average number of nucleotide differences per site between any two sequences	(42–44)
	theta	Nucleotide polymorphism: proportion of nucleotide sites that are expected to be polymorphic in any suitable sample	(45–47)
	hap_diversity_within	Haplotype diversity within the population	(48)
Divergence-based metrics	Divsites	Number of divergent sites	
	K	Nucleotide divergence per base pair, corrected by Jukes-Cantor	(43)
Linkage disequilibrium	Kelly_ZnS	Average pairwise $\underline{r}^2$ value	(49)
	Rozas_ZA	Average of r^2 only between adjacent polymorphic sites	(50)
	Rozas_ZZ	Rozas_ZA minus Kelly_ZnS	(50)
	Wall_B; Wall_Q	Proportion of pairs of adjacent segregating sites that are congruent, with values approaching 1 indicating extensive congruence among adjacent segregating sites	(51)
	iHS	Integrated haplotype score, based on the frequency of alleles in regions of high LD (computed for the autosomes)	(11)
	XP_EHH	Long-range haplotype method to detect recent selective sweeps (computed for the autosomes, between the major continental populations CEU, CHB and YRI, taken in pairs)	(8)
Recombination	recomb_Bherer2017_females/males/sexavg	Recombination estimates (cM/Mb) from the refined genetic map by Bhérer *et al.* (2017), which collects recombination events from six recent studies of human pedigrees, pertaining to a total of 104 246 informative meioses. Maps are available in three separate tracks: females, males and sexavg	(24)
	recomb_deCODE_females/males/sexavg	deCODE genetic map based on 5136 microsatellite markers for 146 families with a total of 1257 meiotic events.	(52)
	recomb_Marshfield_females/males/sexavg	Marshfield genetic map based on 8325 short tandem repeat polymorphisms (STRPs) for 8 CEPH families consisting of 134 individuals with 186 meioses.	(53)
	recomb_Genethon_females/males/sexavg	Genethon genetic map based on 5264 microsatellites for 8 CEPH families consisting of 134 individuals with 186 meioses.	(54)
Selection tests based on SFS and/or variability	FayWu_H	Number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies	(55)
	FuLi_D	Number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants	(29)
	FuLi_F	Number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences	(29)
	Tajima_D	Difference between the number of segregating sites and the average number of nucleotide differences.	(56)
	Zeng_E	Difference between θ_L and θ_W, sensitive to changes in high-frequency variants.	(57)
Selection tests based on the MKT	DoS	Direction of Selection: difference between the proportion of nonsynonymous divergence and nonsynonymous polymorphism	(58)
	NI	Neutrality Index: summarizes the four values in a McDonald and Kreitman test table as a ratio of ratios	(19,59)
	alpha; alpha_cor	Proportion of substitutions that are adaptive. The second is calculated after removing slightly deleterious mutations	(19,32,60,61)

A complete list is available under the section Help → Tracks Description *of PopHuman.*

D1008 Nucleic Acids Research, 2018, Vol. 46, Database issue

Table 3. List of major gene-based variation statistics in PopHuman, computed for each population separately and for different types of sites

Category	Estimate	Description	Reference	Types of sites analyzed
Descriptive statistics	π	Nucleotide diversity: average number of nucleotide differences per site between any two sequences	(42–44)	Whole gene region ±500 bp
	K	Nucleotide divergence per base pair, corrected by Jukes-Cantor	(43)	
	π_a/π_s	Ratio of nonsynonymous to synonymous nucleotide polymorphism (ω)	(44,62)	Ratio: 0-fold divided by 4-fold
	K_a/K_s	Ratio of nonsynonymous to synonymous nucleotide divergence (ω)	(44,62)	
	DAF	Derived Allele Frequency: distribution of allele frequencies of segregating sites	(63)	Whole gene region ±500 bp
Recombination (Bhérer et al. 2017), cM/Mb	cM/Mb	Recombination estimates (cM/Mb) from the refined genetic map by Bhérer *et al.* 2017	(24)	Whole gene region ±500 bp
Standard MKT	P	Number of segregating sites	(42)	Separately: 4-fold; 0-fold; 5′UTR; 3′UTR; intron; intergenic (±500 bp)
	D	Number of divergent sites		
	π	Nucleotide diversity: average number of nucleotide differences per site between any two sequences	(42–44)	
	K	Nucleotide divergence per base pair, corrected by Jukes-Cantor	(43)	
	α	Proportion of substitutions that are adaptive. It is calculated both from P and D, and from π and K	(19,32,60,61)	
Integrative MKT	*d*	Fraction of new mutations that are strongly deleterious and do not segregate in the population	(32)	Separately: 0-fold; 5′UTR; 3′UTR; intron; intergenic (±500 bp)
	b	Fraction of new mutations that are slightly deleterious and segregate at minor allele frequency (MAF) <5%		
	f-γ	Fraction of new mutations that are neutral since before the split of humans and chimpanzees, calculated after removing the excess of sites at MAF <5% due to slightly deleterious mutations		
	γ	Fraction of new mutations that have become neutral recently, after the split of humans and chimpanzees, calculated after removing the excess of sites at MAF <5% due to slightly deleterious mutations		
	α	Proportion of substitutions that are adaptive, calculated after removing slightly deleterious mutations	(19,32,60,61)	
	DoS	Direction of Selection: difference between the proportion of nonsynonymous divergence and nonsynonymous polymorphism	(58)	

A comprehensive explanation is available under the section Help → Integrative MKT of PopHuman.

instance previously developed by our group that hosts population genomics statistics for 30 *Drosophila melanogaster* populations (36). Novel features that have been implemented in PopHuman include the utility to retrieve gene-based integrative MKT metrics.

Compared to the 1000 Genomes Selection Browser 1.0 (18), PopHuman presents three significant advantages. First, PopHuman analyzes the 1000GP Phase III data, which included 2.29 times more sampled sequences (2504 versus 1092) compared to the Phase I, and used an improved variant calling pipeline. Specifically, Phase III implemented an expanded set of variant callers, including some that use haplotype information and others that rely on *de novo* assembly, it considered low coverage and exome sequencing data jointly rather than independently, and used a different genotype calling that allowed the integration of multi-allelic variants and complex events (15). Second, PopHuman analyzes 26 instead of just three populations. This allows detecting very recent selective sweeps that have occurred in a single population and that can only be detected by analyzing data for this specific population; or older selective sweeps shared among a few related populations, whose detection gives a reinforcement of the time depth and biology underlying the specific selection signal. Three illustrative examples are provided: (i) a recent selective sweep related to skin pigmentation (37) in the region comprising the genes *SLC24A5*, *MYEF2*, *SLC12A1* and *CTXN2* in European (EUR) and South Asian (SAS) populations but not in East Asian (EAS) populations (Supplementary Figure S1); (ii) the presence of high frequency derived alleles in the gene *TRPV6* in all non-African populations, with a stronger signature in EAS populations, intermediate in SAS populations, and weaker in EUR populations, reflecting the time frame in which the establishment of agriculture, and thus the corresponding selective sweeps, occurred in those populations (stronger signatures in more recent sweeps; Supplementary Figure S2) and (iii) the presence of high frequency derived alleles in the Duffy red cell antigen gene (*DARC*, *FY*, *ACKR1*) in sub-Saharan Africa, thought to be the result of selection for resistance to *P. vivax* malaria (38,39), which is also seen in EAS populations (Supplementary Figure S3). Finally, PopHuman, contrary to the 1000 Genomes Selection Browser 1.0, implements selection tests based on the comparison of polymorphism and divergence, which are the only ones able to reveal the fixation of adaptive variants and other signatures of recurrent selection occurring over the last millions of years. One extreme example is found in the gene *PRM1*, which encodes a sperm-specific protein that compacts sperm DNA and shows a clear excess of function-altering substitutions between humans and chimpanzees compared to synonymous substitutions, indicative of positive Darwinian selection (40,41) (Supplementary Figure S4).

CONCLUSION

The PopHuman database and browser go a step forward in the description and analysis of the most comprehensive human diversity data to date from a population genomics perspective. We aim PopHuman to be extended to incorporate novel metrics of transcriptomic and epigenomic variation, not only across individuals and species but also during the lifetime of an individual and/or in different parts of the body. In this way, PopHuman will become a pioneer population multi-omics browser advancing the upcoming population –omics synthesis (16).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The authors thank Daniel Rigden and two anonymous referees for helpful comments on the PopHuman implementation and manuscript. We also thank Oscar Conchillo for helpful discussions about the informatics infrastructure in which PopHuman is implemented.

FUNDING

Ministerio de Economía y Competitividad/European Regional Development Fund [grant numbers BFU2013-42649-P to A.B., BFU2016-77961-P to J.B.]; Generalitat de Catalunya [2014-SGR-1346, 2014-SGR-866]; Departament de Genètica i de Microbiologia of the Universitat Autònoma de Barcelona [12ª PIPF to S.H.]; Youth Employment Initiative and European Social Fund [PEJ-2014 to E.S]. Funding for open access charge: Ministerio de Economía y Competitividad/European Regional Development Fund [BFU2013-42649-P to A.B., BFU2016-77961-P to J.B.].

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Consortium,I.H.G.S. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
3. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
4. International HapMap 3 Consortium, Altshuler,D.M., Gibbs,R.A., Peltonen,L., Altshuler,D.M., Gibbs,R.A., Peltonen,L., Dermitzakis,E., Schaffner,S.F., Yu,F. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
5. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
6. International HapMap Consortium, Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
7. Manolio,T.A. and Collins,F.S. (2009) The HapMap and genome-wide association studies in diagnosis and therapy. *Annu. Rev. Med.*, **60**, 443–456.
8. Sabeti,P.C., Varilly,P., Fry,B., Lohmueller,J., Hostetter,E., Cotsapas,C., Xie,X., Byrne,E.H., McCarroll,S.A., Gaudet,R. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.
9. Akey,J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.*, **19**, 711–722.
10. Sabeti,P.C., Reich,D.E., Higgins,J.M., Levine,H.Z., Richter,D.J., Schaffner,S.F., Gabriel,S.B., Platko,J.V., Patterson,N.J., McDonald,G.J. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
11. Voight,B.F., Kudaravalli,S., Wen,X. and Pritchard,J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.
12. Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
13. Goodwin,S., McPherson,J.D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.
14. 1000 Genomes Project Consortium, Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
15. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
16. Casillas,S. and Barbadilla,A. (2017) Molecular population genetics. *Genetics*, **205**, 1003–1035.
17. Sabeti,P.C., Schaffner,S.F., Fry,B., Lohmueller,J., Varilly,P., Shamovsky,O., Palma,A., Mikkelsen,T.S., Altshuler,D. and Lander,E.S. (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.
18. Pybus,M., Dall'Olio,G.M., Luisi,P., Uzkudun,M., Carreño-Torres,A., Pavlidis,P., Laayouni,H., Bertranpetit,J. and Engelken,J. (2014) 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.*, **42**, D903–D909.
19. McDonald,J.H. and Kreitman,M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 652–654.
20. Gazal,S., Sahbatou,M., Babron,M.-C., Génin,E. and Leutenegger,A.-L. (2015) High level of inbreeding in final phase of 1000 Genomes Project. *Sci. Rep.*, **5**, srep17453.
21. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
22. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. *et al.* (2017) Ensembl 2017. *Nucleic Acids Res.*, **45**, D635–D642.
23. Poliakov,A., Foong,J., Brudno,M. and Dubchak,I. (2014) GenomeVISTA–an integrated software package for whole-genome alignment and visualization. *Bioinforma Oxf. Engl.*, **30**, 2654–2655.

24. Bhérer,C., Campbell,C.L. and Auton,A. (2017) Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat. Commun.*, **8**, 14994.
25. Pfeifer,B., Wittelsbürger,U., Ramos-Onsins,S.E. and Lercher,M.J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.*, **31**, 1929–1936.
26. Szpiech,Z.A. and Hernandez,R.D. (2014) selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.*, **31**, 2824–2827.
27. Pybus,M., Luisi,P., Dall'Olio,G.M., Uzkudun,M., Laayouni,H., Bertranpetit,J. and Engelken,J. (2015) Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, **31**, 3946–3952.
28. Haasl,R.J. and Payseur,B.A. (2016) Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol. Ecol.*, **25**, 5–23.
29. Fu,Y.X. and Li,W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.
30. Andolfatto,P. (2005) Adaptive evolution of non-coding DNA in Drosophila. *Nature*, **437**, 1149–1152.
31. Egea,R., Casillas,S. and Barbadilla,A. (2008) Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.*, **36**, W157–W162.
32. Mackay,T.F.C., Richards,S., Stone,E.A., Barbadilla,A., Ayroles,J.F., Zhu,D., Casillas,S., Han,Y., Magwire,M.M., Cridland,J.M. *et al.* (2012) The Drosophila melanogaster Genetic Reference Panel. *Nature*, **482**, 173–178.
33. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
34. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
35. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N., Guruvadoo,L. *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
36. Hervas,S., Sanz,E., Casillas,S., Pool,J.E. and Barbadilla,A. (2017) PopFly: the Drosophila population genomics browser. *Bioinformatics*, **33**, 2779–2780.
37. Lamason,R.L., Mohideen,M.-A.P.K., Mest,J.R., Wong,A.C., Norton,H.L., Aros,M.C., Jurynec,M.J., Mao,X., Humphreville,V.R., Humbert,J.E. *et al.* (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, **310**, 1782–1786.
38. Escalante,A.A., Cornejo,O.E., Freeland,D.E., Poe,A.C., Durrego,E., Collins,W.E. and Lal,A.A. (2005) A monkey's tale: the origin of Plasmodium vivax as a human malaria parasite. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 1980–1985.
39. Hamblin,M.T., Thompson,E.E. and Di Rienzo,A. (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.*, **70**, 369–383.
40. Wyckoff,G.J., Wang,W. and Wu,C.I. (2000) Rapid evolution of male reproductive genes in the descent of man. *Nature*, **403**, 304–309.
41. Rooney,A.P. and Zhang,J. (1999) Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? *Mol. Biol. Evol.*, **16**, 706–710.
42. Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, NY.
43. Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In: *Mammalian Protein Metabolism*. Academic Press, NY, pp. 21–32.
44. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
45. Watterson,G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**, 256–276.
46. Tajima,F. (1993) Mesurement of DNA polymorphism. In: *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*. Sinauer Associates Inc., Suderland, Massachusetts.
47. Tajima,F. (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, **143**, 1457–1465.
48. Hudson,R.R., Slatkin,M. and Maddison,W.P. (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.
49. Kelly,J.K. (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.
50. Rozas,J., Gullaud,M., Blandin,G. and Aguadé,M. (2001) DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. *Genetics*, **158**, 1147–1155.
51. Wall,J.D. (1999) Recombination and the power of statistical tests of neutrality. *Genet Res*, **74**, 65–79.
52. Kong,A., Gudbjartsson,D.F., Sainz,J., Jonsdottir,G.M., Gudjonsson,S.A., Richardsson,B., Sigurdardottir,S., Barnard,J., Hallbeck,B., Masson,G. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–247.
53. Broman,K.W., Murray,J.C., Sheffield,V.C., White,R.L. and Weber,J.L. (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.*, **63**, 861–869.
54. Dib,C., Fauré,S., Fizames,C., Samson,D., Drouot,N., Vignal,A., Millasseau,P., Marc,S., Kazan,J., Seboun,E. *et al.* (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, 152–154.
55. Fay,J.C. and Wu,C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.
56. Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
57. Zeng,K., Fu,Y.-X., Shi,S. and Wu,C.-I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, **174**, 1431–1439.
58. Stoletzki,N. and Eyre-Walker,A. (2011) Estimation of the Neutrality Index. *Mol. Biol. Evol.*, **28**, 63–70.
59. Rand,D.M. and Kann,L.M. (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Mol. Biol. Evol.*, **13**, 735–748.
60. Charlesworth,B. (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res.*, **63**, 213–227.
61. Smith,N.G. and Eyre-Walker,A. (2002) Adaptive protein evolution in Drosophila. *Nature*, **415**, 1022–1024.
62. Li,W.H., Wu,C.I. and Luo,C.C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, **2**, 150–174.
63. Ronen,R., Udpa,N., Halperin,E. and Bafna,V. (2013) Learning natural selection from the site frequency spectrum. *Genetics*, **195**, 181–193.

List of Tables

List of Figures

List of Boxes

List of Supplementary Tables

List of Supplementary Figures