

# Visualization, description and analysis of the genome variation of a natural population of *Drosophila melanogaster*

*Visualització, descripció i anàlisi de la variació genòmica d'una població natural de Drosophila melanogaster*

*Visualización, descripción y análisis de la variación genómica de una población natural de Drosophila melanogaster*

**Doctoral Thesis**  
PhD. in Genetics

Miquel Ràmia Jesús

Director: Antonio Barbadilla

**UAB**  
Universitat Autònoma  
de Barcelona

Departament de Genètica i de Microbiologia  
Facultat de Biociències  
Universitat Autònoma de Barcelona  
Bellaterra, 2015



**“Art is I; science is we”**  
(Claude Bernard)

---

# INDEX

Preface	8
Part 1. Introduction	12
1.1 Evolution and Population genetics	12
1.1.1 Molecular population genetics and the Neutral Theory	14
1.1.2 Mutation as the ultimate source of genetic variation	16
1.1.3 The population dynamics of genetic variation	20
1.1.4 Explaining genome-wide patterns of diversity	21
1.1.5 The estimation of DNA variation	26
1.1.6 Detecting natural selection in the genome-wide	27
1.1.7 Tests of selection	28
1.1.8 Detecting selection genome-wide	33
1.2 Population genomics: population genetics meets genomics	34
1.2.1 The Drosophila Genetic Reference Panel	35
1.3 Genome Browsers	38
1.3.1 Molecular Databases	38
1.3.2 Genome Browsers	41
1.3.3 GMOD community and the Generic Genome Browser	43
1.4 Objectives	45
Part 2. Materials and Methods	48
2.1 DGRP Input data	48
2.1.1 Sequence data	49
2.1.2 Recombination data	50
2.1.3 Diversity measures	51
2.2 PopDrowser: the Population Drosophila Browser	51
2.2.1 Selection of GBrowse as a framework for the PopDrowser	51

2.2.2 Interface and implementation	52
2.3 Nucleotide variation description and analysis along the genome of a natural population of <i>Drosophila melanogaster</i>	57
2.3.1 Diversity measures & Linkage disequilibrium	57
2.3.2 Re-coded whole-genome consensus sequence	58
2.3.3 Detecting and estimating natural selection	58
2.4 Indel variation landscape in the genome of a natural population of <i>D. melanogaster</i>	61
2.4.1 Filtering structural variants	61
2.4.2 Inferring the ancestral state of Indels	62
2.4.3 Calculating Indel Variation	63
2.4.4 Estimating the proportion of adaptive fixations in indels	64
2.5 Recapitulation of used and developed software	65
Part 3. Results	68
3.1 PopDrowser: the <i>Drosophila</i> Genome Variation Browser	68
3.1.1 Output	68
3.1.2 Custom analyses on-the-fly	71
3.2 Nucleic variation analysis of a natural population of <i>Drosophila melanogaster</i>	72
3.2.1 Polymorphism and Divergence in the chromosome arms of <i>D. melanogaster</i>	72
3.2.2 Recombination landscape	73
3.2.3 Mapping selection across the genome	76
3.3 Indel variation landscape in the genome of a natural population of <i>D. melanogaster</i>	80
3.3.1 Genome-wide indel statistics and distribution	80
3.3.2 Covariation between SNP and indel variation	83
3.3.4 Derived allele frequency spectrum	85
3.3.5 Adaptive selection on indel variation	87

Part 4. Discussion	92
4.1 Population genomics software development	92
4.1.1 PopDrowser and Genome Brower	92
4.1.2 Integrative-MKT implementation	95
4.2 Nucleic Variation in a natural population of <i>D. melanogaster</i>	97
4.2.1 Genome-wide patterns of polymorphism and divergence and recombination	97
4.2.2 Mapping natural selection across the genome and the major effect of recombination	100
4.2.3 Fast-X hypothesis	102
4.3 Indel Variation Landscape in <i>D. melanogaster</i>	105
4.3.1 Genome-wide description of indel variation and indel-SNP relationships in <i>D. melanogaster</i>	105
4.3.2 Selection acting on indels	107
Part 5. Conclusions	112
Bibliography	115
Annex A. PopDrowser: the Population Drosophila Browser	128
Annex B. The <i>Drosophila melanogaster</i> genetic reference panel	132
Annex C. Natural variation in genome architecture among 205 <i>Drosophila melanogaster</i> genetic reference panel lines	140
Acknowledgements	157

---

## Preface

The description and explanation of genetic variation within and between populations, the goal of population genetics since its origins, have been hampered by decades because of the technical inability to directly measure the genetic variation of populations. The present genome era, with the explosive growth of genome sequences fueled by the next-generation sequencing technologies, has led us to the present golden age of the study of genetic variation at the genome scale. Population genetics is no longer an empirically insufficient science, but it is more than ever a research field where bioinformatics tools for data mining and management of large-scale dataset, statistical and evolutionary models, and advanced molecular techniques of mass generation of sequences are all them integrated in an interdisciplinary endeavor. As a consequence of this breakthrough, a new 'omic' discipline has emerged: Population Genomics.

But, what is Population Genomics? For Charlesworth (2010), it's simply "a new term for a field of study as old as Genetics itself". It's the 'old field' of *Population Genetics* when studying the amount and causes of variability in natural populations in a genome-wide fashion.

This thesis is both a population genomics study and a bioinformatics project centred on the visualization, description and analysis of the genome-wide DNA variation data from a natural population of model organism *Drosophila melanogaster*. The data used has been obtained by the international initiative The Drosophila Genetic Reference Panel (DGRP) (Mackay *et al.* 2012). DGRP has sequenced the complete genomes of 158 (freeze 1) and 205 (freeze 2) inbred lines of *Drosophila melanogaster* from a single natural population of Raleigh (USA). A major goal of this project was to create a resource of common genetic polymorphism data to further perform population genomics analyses.

The DGRP sequence data has allowed us to carry out a thorough study of genome-wide variation in a natural population of *D. melanogaster*. After developing a complete, public and accessible map of the polymorphism present in this population, we have described the patterns of polymorphism and divergence (nucleotide and indel variants) along chromosome arms. We observe a clear and consistent pattern of genome nucleotide diversity along arms of the autosomic chromosomes both for SNP and indels: nucleotide diversity is reduced on average in centromeric regions relative to non-centromeric regions, and at the telomeres. This pattern is not observed in the X chromosome, where diversity is almost uniform all along the chromosome. Polymorphism and recombination are correlated along chromosome arms, but only for those regions where recombination rate is below  $2\text{cM Mb}^{-1}$ . Recombination rate seems to be the



major force shaping the patterns of polymorphism along chromosome arms and its effect seems to be mediated by its impact on linked selection.

We have mapped the footprint of natural selection on SNP and indel variants throughout the genome, observing a pervasive action of natural selection, both adaptive and purifying selection. Adaptive selection occurs preferentially in non-centromeric regions. Natural selection acts differently between insertions and deletions, being deletions more strongly selected by purifying selection, which supports the mutational equilibrium theory for genome size evolution.



---

## **Part 1.**

### **Introduction**

---

# 1. Introduction

## 1.1 Evolution and Population Genetics

---

Since Charles Darwin most universally known publication, *The origin of Species* (1859), biological evolution is understood as a population process, where phenotypic and genotypic 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 on Earth (Lewontin 1974). Biological evolution is the result of this elementary process of change in populations through generations.

Formally, evolution occurs if two conditions are met: (i) variation in phenotypic traits within a population, and (ii) inheritance of this variation, in other words, variation must be heritable at least partially among generations (Lewontin 1970; Endler 1986). DNA is the molecule that carries the genetic information (Avery *et al.* 1944), and among its properties two are essential to the evolutionary process. On one hand, the molecule is intrinsically mutable, being this the origin of genetic variation. On the other hand, it allows the replication of old and new variants from one generation to another. The reproductive or survival advantage or disadvantage an individual has for carrying a given variant relative to individuals that does not have it is called *fitness*. Only when genetic variants provide individuals with differences in fitness, the process of *natural selection* described by Darwin can occur (Endler 1986). In consequence, the action of natural selection implies evolution (except in the case of balancing selection, where 2 or more variants are maintained without change among generations), but natural selection is not a necessary condition for evolution to occur.

Within the variation paradigm, population genetics provides the theoretical framework to describe how biological evolution does occur. The main aim of population genetics is the

description and interpretation of genetic variation within and among populations (Dobzhansky 1937). The Hardy-Weinberg principle, the single mathematical model formulated independently by G. H. Hardy and W. R. Weinberg in 1908 served as a null model to explain the maintenance of genetic variation in a population during the first years of genetics. The principle states that in an ideal population and in absence of any other evolutionary forces, allele frequencies would remain unchanged generation after generation. Population genetics is conceived as a theory of forces that can affect allele frequencies in a population. These forces are principally *mutation, migration, natural selection, recombination* and *random genetic drift*.

The mathematical foundations of population genetics were established by R. A. Fisher, J. B. S. Haldane and S. Wright in the second and third decades of the XX century. They figured out the consequences of chance and selection in populations with Mendelian inheritance, and turned population genetics into the explanatory core of the evolutionary theory. In the late 1930s and 40s, the integration of theoretical population genetics with other evolutionary research fields such as experimental population biology, palaeontology, systematics, zoology and botany gave rise to the *Modern Synthesis* of evolutionary biology (Dobzhansky 1937; Mayr 1942; Simpson 1944; Stebbins 1950). The main difference between the modern synthetic theory and Darwin's original view of evolution by natural selection is the addition of the Mendelian laws of heredity in a population genetics framework. This new theory is also called Neo-Darwinism by some, although the term was coined years before by George Romanes referring to the theory of Alfred Russel Wallace and August Weismann to differentiate it from the initial Darwin's theory (Romanes 1906).

The Modern Synthesis theory considers natural selection the most fundamental process underlying evolution in detriment of drift and other non-adaptive forces. In a first attempt to account for the nature of genetic variation, two different models were put forward (Lewontin 1974). The *classical model* supported the role of natural selection as purging populations of new mutations and thus predicted that most gene loci are homozygous for the wild-type allele (Muller and Kaplan 1966). On the other hand, the *balance model* considers that natural selection maintains high levels of genetic diversity in populations by favouring heterozygosity at many gene loci (Dobzhansky 1970; Ford 1971). The balance model could account for why a population can respond quickly to environmental changes by selecting variation already existing in the population and changing its frequencies. This debate moved to a more subtle one after the first estimation of genetic diversity using gel electrophoresis techniques, and the first descriptions of protein allelic variants.

### 1.1.1 Molecular population genetics and the Neutral Theory

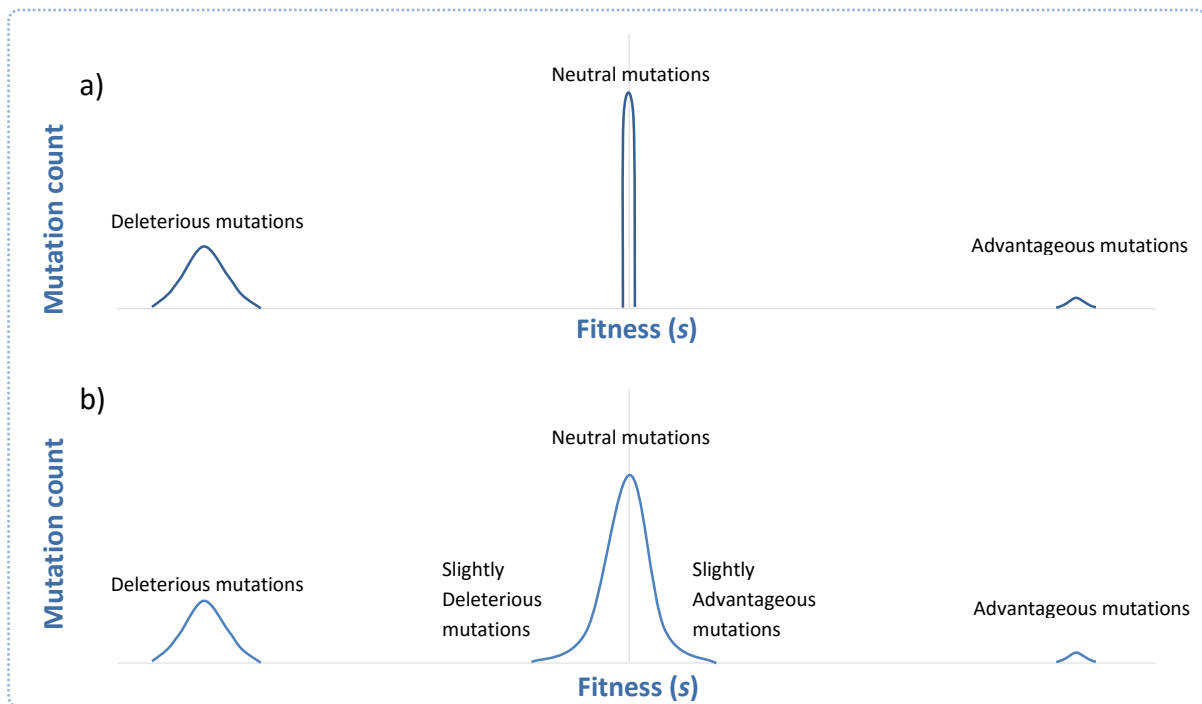
With the advent of the electrophoretic techniques to estimate protein variation, population genetics entered in the molecular age, the so-called 'Allozyme era' (Lewontin 1974; 1992). The results of electrophoretic experiments exposed substantial amounts of genetic variation in most populations (Nevo *et al.* 1984), much more than expected, and seemed to better support the balance model than the classical model. Levels of genetic diversity were also found to vary in a non-random way among populations, species, higher taxa and several ecological, demographic and life history parameters (Nevo *et al.* 1984).

At the time, a new theory was put forward to explain the patterns of molecular genetic variation within and among species, in a complete opposite way than the balance hypothesis, selective based, does. Kimura's *Neutral Theory of molecular evolution* states that most of new mutations at the molecular level are either strongly deleterious or selectively neutral, and therefore the dynamics of polymorphism in populations are determined by random genetic drift rather than by natural selection (Kimura 1968, Kimura 1983). Some of the principal implications of the neutral theory are:

1. Deleterious mutations are rapidly removed from the population, and adaptive mutations are rapidly fixed; therefore, most variation within species is selectively neutral (Figure 1.1).
2. A steady-state rate at which neutral mutations are fixed in a population ( $k$ ) equals the neutral mutation rate:  $k = \mu_0$ , where  $\mu_0$  is the neutral mutation rate,  $\mu_0 = f_{neutral} \mu$ , where  $f_{neutral}$  is the proportion of all mutations that are neutral and  $\mu$  the intrinsic mutation rate by generation. If all mutation are neutral, then  $\mu_0 = \mu$ .
3. The level of polymorphism in a population ( $\vartheta$ ) is a function of the neutral mutation rate and the effective population size ( $N_e$ ):  $\vartheta = 4N_e \mu_0$ .
4. Polymorphisms are transient (on their way to loss or fixation) rather than balanced by selection. Larger populations are expected to have a higher heterozygosity, as reflected in the greater number of alleles segregating at a time.

The hypothesis of selective neutrality would also apply to most nucleotide or amino acid substitutions that occur during the course of evolution. Still, Kimura emphasized the compatibility of his theory, mainly based in mutation and drift at the molecular level, with natural selection shaping variation at the phenotypic/morphological level. There have been new

refinements to the neutral theory, especially the nearly-neutral and slightly deleterious mutation hypotheses of Tomoko Ohta (Ohta 1995), that modifies the original theory considering that slightly deleterious variants could still segregate at low frequencies in the population (Figure 1.1 and Box 1). In any case, Kimura's neutral theory became the theoretical foundation of molecular population genetics.



**Figure 1.1** Representation of the Distribution of Fitness Effects (DFE) for mutations under the Kimura's neutral theory (a) where mutations are considered to be only neutral, advantageous or deleterious; and Ohta's nearly neutral theory (b) where is considered that some mutations are not completely neutral but either slightly advantageous or slightly deleterious. 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.

A consequence of the neutral theory is the existence of a *random molecular clock*, previously inferred from protein sequence data by Zuckerkandl and Pauling (1962). Assuming that the neutral mutation rate is equal to the neutral allele fixation rate ( $k = \mu_0$ ), when two populations or species split, the number of genetic differences among them is proportional to the time of speciation. This can be used as a molecular clock since the number of differences among sequences from different species represents the relative times of divergence among them. Related to this idea of a molecular clock, the *Coalescent Theory* (Kingman 2000) tries to trace the changes suffered by a genomic region, shared by different members of a population or

different taxa, to a single ancestral copy: the most recent common ancestor (MRCA). The mathematical methods created around this theory allowed the construction of coalescent graphs, gene genealogies similar to phylogenetic trees, which try to describe the phylogenetic and genealogical relationships between the different sequences.

### 1.1.2 Mutation as the ultimate source of genetic variation

*Genetic variation* is the cornerstone of the evolutionary process. Heritable variation in any trait must exist before it can undergo any process of adaptation by natural selection. Hence, the study of variation within individuals and populations is crucial to understand every process of evolutionary change. But for many years during the 19<sup>th</sup> century and the start of the 20<sup>th</sup> century, variation could only be studied for phenotypic traits, where discrete Mendelian variation is rare and quantitative traits are abundant. A phenotypic trait results from the interaction between a given genotype (which is heritable) and a specific environment. Observed phenotypes are the final result of many interactions difficult to discern.

A *mutation* is an adaptively non-directed change in the genomic sequence of an individual, and mutations in the DNA molecule are the ultimate source of genetic variation. Once a new variant appears by mutation in the DNA it can be replicated and transmitted from generation to generation. Gel electrophoresis of proteins assesses indirectly genetic variation (Johnson *et al.* 1966; Lewontin and Hubby 1966, Harris 1966). It was not until the late 70's that actual variation in the DNA molecule was analysed using first restriction enzymes (Avice *et al.* 1983), and later, with the milestone of *sequencing technologies* (Sanger and Coulson 1975, Maxam and Gilbert 1977), genetic variation was estimated at the ultimate DNA sequence level (Kreitman 1983). The automation and parallelization of the Sanger method was the key that provided us with an impressive number of sequenced genomes in practically 20 years. Nowadays, more advanced and high throughput second or next generation sequencing (NGS) methods are used to analyse several types of variation in the DNA sequence, and with third generation methods at hand, even more advances are to be expected (Niedringhaus *et al.* 2011, McGinn & Gut 2012)

Mutation size in genomes ranges from single nucleotide changes to microscopically visible karyotypic alterations, where they can be, for instance, larger than 3Mb in humans (Feuk *et al.* 2006, Conrad and Hurles 2007). Accordingly, mutations are categorized in two non-overlapping types: (I) single-nucleotide variants or *single nucleotide polymorphisms* (SNPs) when only one nucleotide in the genome is mutated, and (II) *structural variation*, when multiple bases are



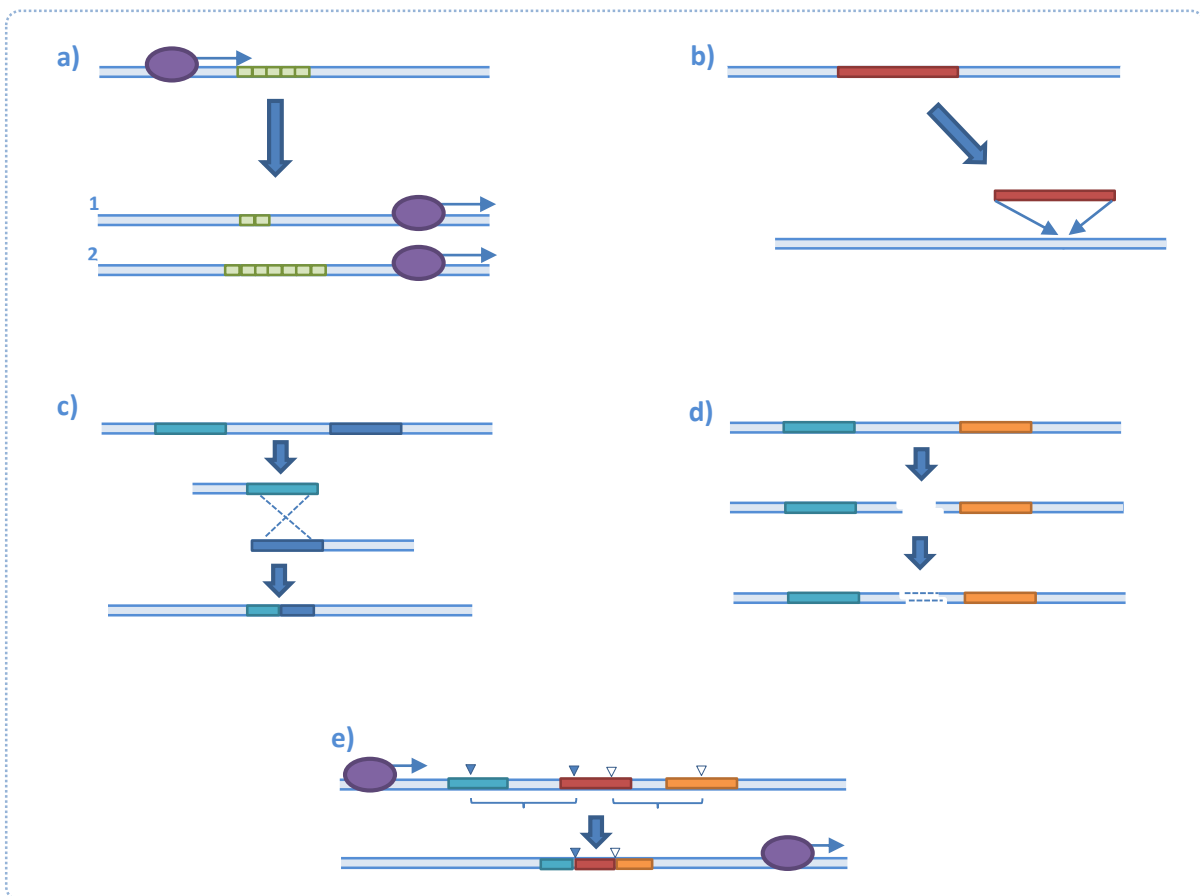
involved. Structural variants in the genome comprise insertions/deletions (indels), inversions, translocations and variations in the number of copies of a given genomic segment (Table 1.1). The term *structural variation* is commonly defined as variation of more than one nucleotide. But this could lead to confusion, since in the literature structural variation is also defined as that involving 'segments of DNA longer than 1kb' (Feuk *et al.* 2006).

The first studies of DNA variation focused on single-nucleotide differences among individuals. Although only one nucleotide is affected, their abundance in the genome makes them the most frequent source of inter-individual genetic variation event with respect to other variation types. Single nucleotide polymorphisms were believed to account for >90% of the genomic variability in humans (Collins *et al.* 1998).

Regardless of their minor amount when compared to the number of single-base variants, all these others structural variants comprise a significant fraction of a genome since each one involve longer segment of DNA than SNPs. In the case of copy number variations (CNVs), studies show they represent a range from 3.7% to 12% (112.7 - 360Mb) of the human genome (Redon *et al.* 2006, Conrad *et al.* 2010) with more recent studies defining a more precise range of 4.8% to 9.5% of CNV contribution to the genome (Zarrei *et al.* 2015). The combination of all the structural variants in a genome implies seemingly more DNA in play than the DNA assembled by single-nucleotide variants, taking into account that the estimated number of SNPs reported for the human genome are 149,735,377 (dbSNP, June 2014). Following SNPs, indels are the next most abundant form of genetic variation and are the most common type of structural variants (Väli *et al.* 2008, Mullaney *et al.* 2010), specially short ( $\leq 50$ bp) insertion and deletions (indels), at least when looking at the human genome (Montgomery *et al.* 2013).

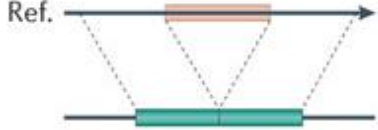
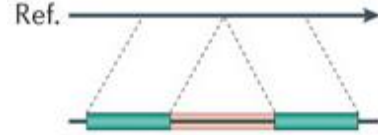
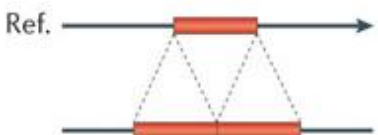

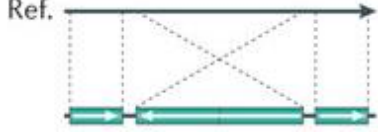

Indels and structural mutations can have various mechanisms of origin. Overall, formation mechanisms differ for the size of the variant. According to Pang *et al.* (2013) in a human genome, small variants (< 1Kb), are associated with nonhomologous processes 72.6% of the time, in contrast with 24.9% of microsatellite events. Medium size variants (<10Kb) are commonly related to minisatellites (25.8%) or retrotransposons (24%) among other causes. Finally, a 46.2% large variants (>10%) seem to be associated with nonallelic homologous recombination (Pang *et al.* 2013).

Looking specifically on indel formation mechanisms, one of the most studied is the polymerase slippage (Streisinger *et al.* 1966, Levinson and Gutman 1987, Greenblatt *et al.* 1996, Taylor *et al.* 2004, Montgomery *et al.* 2013). However, indels can also be originated by other mechanisms, that may be responsible for other structural variants as well, like imperfect repairs of double-strand breaks (Chu 1997, McVey *et al.* 2004), fork stalling and template switching (FoSTeS), microhomology-mediated break-induced replication (MMBIR) (Lee *et al.* 2007, Hastings *et al.* 2009) and hairpin loop formation due to presence of palindromic sequences (Greenblatt *et al.* 1996, Hastings *et al.* 2009) (Figure 1.2).



**Figure 1.2** Representation of some mechanisms of indel and structural variants formation. (a) **Polymerase slippage** (or slipped strand mispairing) is a mutation process where regions of small repeats can be expanded or contracted by action of the polymerase complex during replication. (b) **Transposable elements** can cut or copy fragments of DNA and insert them in other locations of the genome. In (c) the two blue coloured fragments share high homology, which align in a **non-allelic homologous recombination** (NAHR) event and can produce deletion or duplication of part of the homolog fragments and flanking regions. (d) A **non-homologous end-joining** (NHEJ) event. A double strand break occurs between the blue and orange non-homologous fragments. The NHEJ mechanism modifies and re-joins the ends resulting in a deletion between the two fragments. Finally (e) a **fork stalling and template switching** (FoSTeS) event, where fragments between microhomology segments (2 to 5 bp, represented as triangles) can be deleted during the replication. [c,d and e adapted from Gu *et al.* 2008]

**Table 1.1** Common DNA mutation types.

Type of variation	Description	
1. Single nucleotide polymorphisms (SNP)	Base substitution involving only a single nucleotide. It can be transitions or transversions. Coding-related mutations can be missense, nonsense, silent or splice-site mutations.	<p>ATGCAGTCGATCG<b>A</b>TGGCATGCATGC            ATGCAGTCGATCG<b>C</b>TGGCATGCATGC</p>
2. Insertions and deletions (Indel)	Extra base pairs that may be added (insertions) or removed (deletions) from the DNA.	<p><i>Deletion:</i></p>  <p><i>Insertion:</i></p> 
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.	
4. Copy number variations (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. Segmental duplications	Specific case of CNV where a pair of DNA fragments $>1$ kb share $>90\%$ identity	
6. Inversions	Change in the orientation of a piece of a DNA segment.	
7. Translocations	Transfer of a piece of a chromosome to a nonhomologous chromosome. It can often be reciprocal.	

[Adapted from Casillas 2007, Freeman *et al.* 2006 and Alkan 2011]

### 1.1.3 The population dynamics of genetic variation

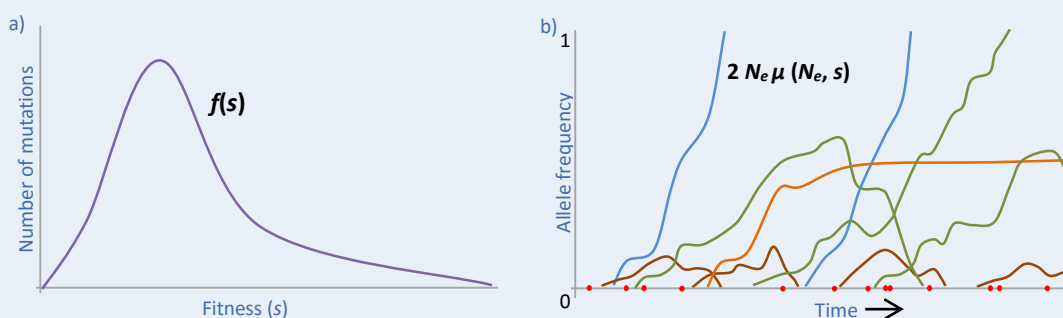
Which fraction of new mutations is deleterious, neutral, or advantageous? This question has been part of the debate since population genetics started its path into the molecular era. However, classifying mutations into these groups may not reflect correctly the real fitness effect of mutations. In reality, mutations have a continuous *Distribution of Fitness Effects* (DFE) ranging from lethal or very deleterious, through slightly deleterious, neutral, slightly advantageous and strongly advantageous (Keightley & Eyre-Walker 2010, Piganeau and Eyre-Walker 2003) (Figure 1.1). Ultimately, the levels of genetic variation observable in a given genome region in a population is a combination of their DFE of new mutations and their population dynamics over time (see Box 1).

#### Box 1: Two main functions determining polymorphism and divergence in a population

The *distribution of fitness effects* (DFE) can be described as the relative frequency of mutations that range from deleterious, through neutral, to advantageous contribution in the population. The DFE can be mathematically defined as a function of the fitness (measured by the coefficient of selection,  $s$ ) of new mutations entering in the population:  $f(s)$  (Figure 1.3a).

However, the observable level of genetic variation in a population is also affected by the population dynamics of each mutation from the moment it appears through time; this dynamics is mainly defined by the probability of fixation of each mutation once it appears, which depends on the effective population size ( $N_e$ ) and the fitness ( $s$ ):  $2N_e\mu(N_e, s)$  (Figure 1.3b). Neutral alleles reach fixation or disappear from the population by random mechanisms. Advantageous variants become fixed quickly and, contrarily, slightly deleterious mutations segregate to some extent until they are removed. Interestingly, alleles under balancing selection tend to stabilize at an intermediate frequency in the population never reaching fixation. The most extreme case is strongly deleterious mutations which are never observed as polymorphism due to their fast elimination from the population.

Finally, we can calculate divergence ( $k$ ) as the integral or weighted sum of the combined probability of fixation and fitness effect, from fitness  $-\infty$  to  $\infty$ :  $\int_{-\infty}^{\infty} 2N_e\mu(N_e, s) f(s) ds$ .



**Figure 1.3** Factors determining the substitution rate of new mutations in populations. (a) Distribution of fitness effects (DFE) and (b) a diagram showing the dynamics of different types of alleles in a population through time. In (b), 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). In green are represented the dynamics of neutral alleles, in blue are advantageous alleles, in brown are slightly deleterious alleles, orange is for alleles under balancing selection, and red dots represent strongly deleterious mutations. [Adapted from Hartl and Clark (1997)]

Once a mutation appears in a population, its frequency starts a journey whose fate is determined by population genetics factors. Most mutations will be lost from the population in the same generation they appear either by chance or because the individual carrying it dies before leaving offspring. However, a mutation could increase its frequency in the population through generations, either by random genetic drift or because it gives some advantage to the individuals that possess it. The state in which multiple alleles exist for a same locus within the population is called *polymorphism*. If sometime in the future a single allele is shared by all the individuals within a population we say that this allele has reached *fixation*. If a newly appeared allele is neutral, the probability that it becomes eventually fixed is its initial frequency. For a new mutation present in a single individual this probability is  $1/2N$  for diploid or  $1/N$  for haploid organisms. This chance of fixation is influenced for the above mentioned population genetics forces, like the fitness effect of the allele (see Box 1).

The accumulation of distinct allele fixations between two different populations is referred to as *divergence*. The independent allele fixation along two populations of the same species which are reproductively isolated for many generations can derive in two new different species (*speciation*). The ultimate consequence of this continuous process is the rich diversity in life forms we can see in Earth.

Polymorphism and divergence tell us different and complementary stories about the past and present events of a population. When we analyse the polymorphism in a population, we are actually observing a kind of snapshot of the variation dynamics at that precise moment in time, and it also allows us to infer events that have happened recently. On the other hand, when studying divergence between species, we are observing (putatively) fixations between them, a process that takes a longer time and tells us about more ancient events. The combined analysis of polymorphism and divergence is one of the most powerful approaches to understand the influence of different population genetics forces modelling the patterns of molecular evolutionary change.

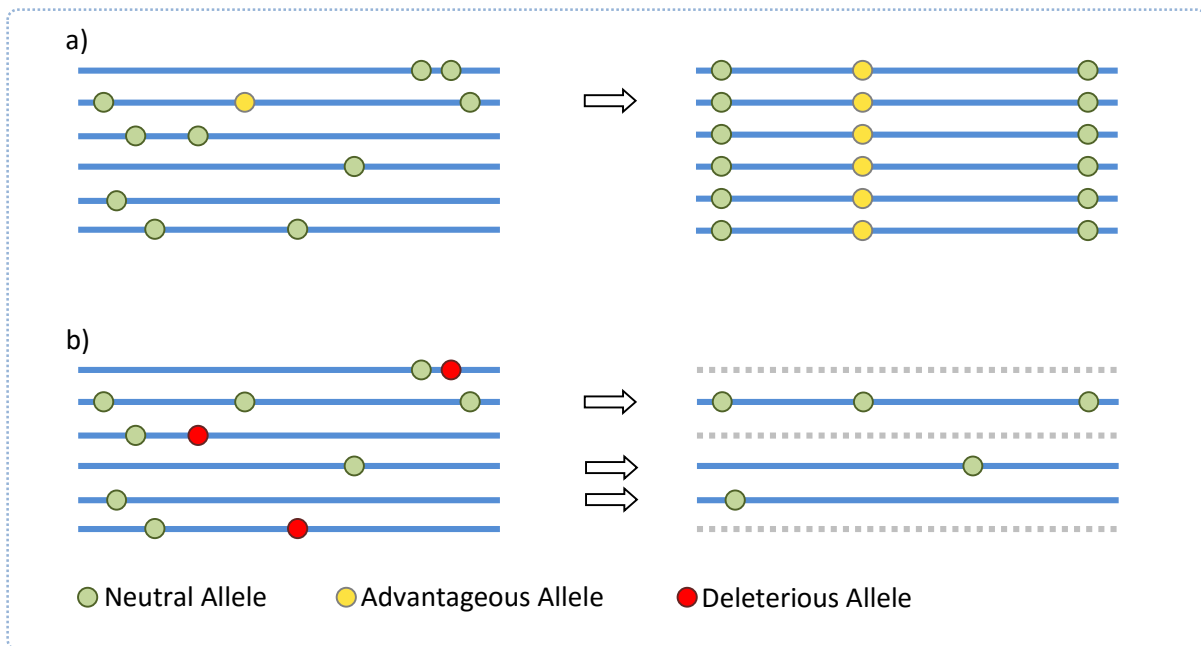
#### 1.1.4 Explaining genome-wide patterns of diversity

Even with all the available data and techniques to study genetic diversity, it's still not completely understood how different evolutionary forces contribute to the patterns of genetic variation we observe nowadays. The forces shaping the genetic structure of populations tend to be weak,

and also they take action very slowly during thousands or even millions of years, which makes any observation difficult to decipher.

**Genetic drift.** If we take into account the neutral theory, the main two forces affecting genetic variation are *mutation* and *random genetic drift*. Mutation adds new variation to a population at the rate  $2N\mu$  in diploid organisms (where  $N$  is the population size). Drift instead, removes variation from the population at each generation at a rate depending on the population size ( $1/2N_e$ ). This implies that in small populations drift removes variation faster than new variation is added by mutation. On the contrary, on large populations drift is not strong enough to remove all the variation that appears steadily. If genetic drift is the determinant force, a lineal relationship between diversity and population size would be expected: the larger population sizes the larger genetic variation. From Kimura's neutral theory, the total number of mutations that will be fixed at a given gene or DNA region each generation is  $(2N\mu)(1/2N)$ , which is the already mentioned mutation rate of the DNA region, and the probability that this mutation will be fixed. However, it takes time for a new mutation to achieve fixation once it appears in the population, this time depends on the population size and is  $4N_e$ ; thus, in a mutation-drift equilibrium, the average number of polymorphic sites when comparing two random sequences within the same population is  $4N\mu$  (Kimura 1983), this is also known as the neutral population mutation rate  $\theta$  or the expected neutral nucleotide heterozygosity (see 1.1.6).

**The paradox of variation and linked selection theories.** However, the first studies of allozyme polymorphisms did not completely reflect the supposed lineal proportion between population size and genetic diversity, a phenomenon that has been called 'The paradox of variation' (Lewontin 1974, Hahn 2008). **Genetic hitchhiking** was proposed as an explanation to some lower than expected levels of diversity (Smith and High 1974; Kaplan, Hudson and Langley 1989). In this process, neutral alleles near a favourable mutation can go together to fixation (also called *selective sweep*), resulting in reduced variation in a region. Since variants linked to a selected site are also affected by selection, the region is undergoing **linked selection** (Figure 1.4). Later, Gillespie (2000 a and b) refined the concept taking into account both the effects of neutral theory's genetic drift and repetitive genetic hitchhiking, what it is called **Genetic Draft** (Gillespie 2000a; Gillespie 2000b; Gillespie 2001, Sella *et al.* 2009). Generally, genetic drift removes variation from the population, but in a population large enough there is a possibility of having recurrent hitchhiking events. In this scenario, genetic variation tends to increase, and also the frequency of hitchhiking events which reduce genetic diversity as well.

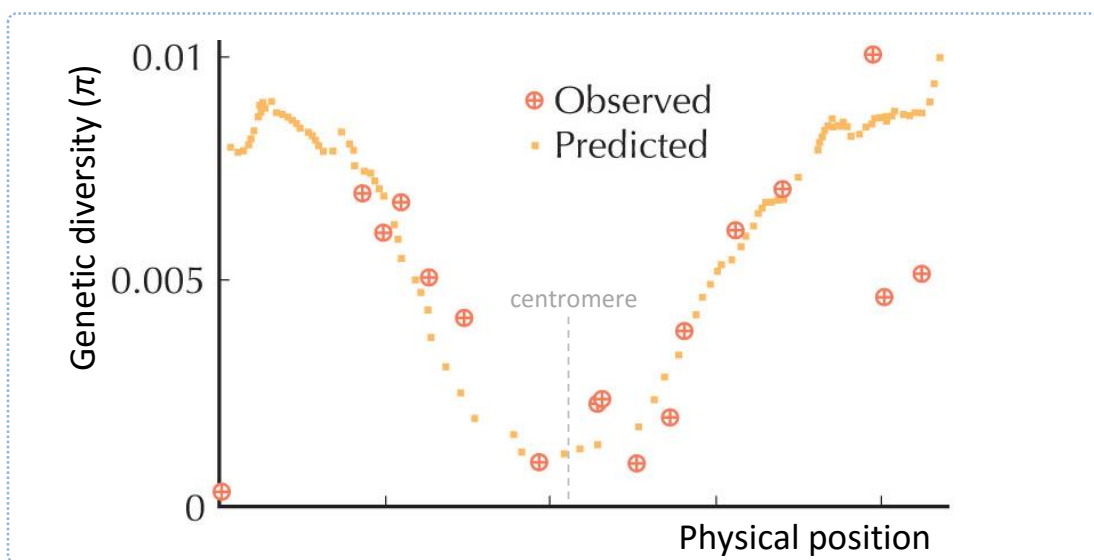


**Figure 1.4** Hitchhiking and Background selection effects on genome variation. Hitchhiking of neutral alleles linked to an advantageous allele being selected (a) results in a reduction in variation and the homogenization of the region before recombination breaks it down over time. Background selection (b) happens when whole regions are selected against due to deleterious alleles. This results in a heterogeneous reduction of genetic diversity with presence of rare alleles.

Linked selection can also occur with negative or purifying selection. **Background selection** is the process in which non-deleterious diversity is removed from the population for being linked to deleterious sites. The effect in this case is a reduction of the number of chromosomes that contributes to the next generation, which is formally identical to that of a reduction in population size except that the reduction applies, not to the genome as a whole, but to a tightly linked region (Charlesworth *et al.* 1993) (Figure 1.4).

**Recombination as a key factor mediating the fate of linked sites in the genome.** A major process that must be considered to explain patterns of genome-wide diversity is recombination. When the first DNA variation analyses appeared in the 80's, lower levels of variation were observed in *Drosophila* in regions of low recombination such as near the centromeres (Aguade *et al.* 1989; Stephan and Langley 1989; Berry *et al.* 1991; Begun and Aquadro 1992; Martin-Campos *et al.* 1992; Stephan and Mitchell 1992; Langley *et al.* 1993) (Figure 1.5). One first explanation was that recombination is itself mutagenic (or both mutation and recombination have common mechanisms). However, patterns of divergence did not seem to increment in high recombination regions as would be expected according neutral theory for larger mutational regions. Hence, increased mutation rate associated with recombination does not seem the

explanation for correlation between recombination and polymorphism. Instead, linked selection events, such as positive selective sweeps or negative background selection, could produce this effect, since loci in high recombination regions are more prone to escape from the effects of selection on nearby sites (Begun and Aquadro 1992). Moreover, correlation between recombination and divergence is not expected under these models. Birky and Walsh (1988) demonstrated that linked selection has no effect on long term neutral fixation, so a linked selection event would reduce polymorphism levels with no effect in the divergence levels.



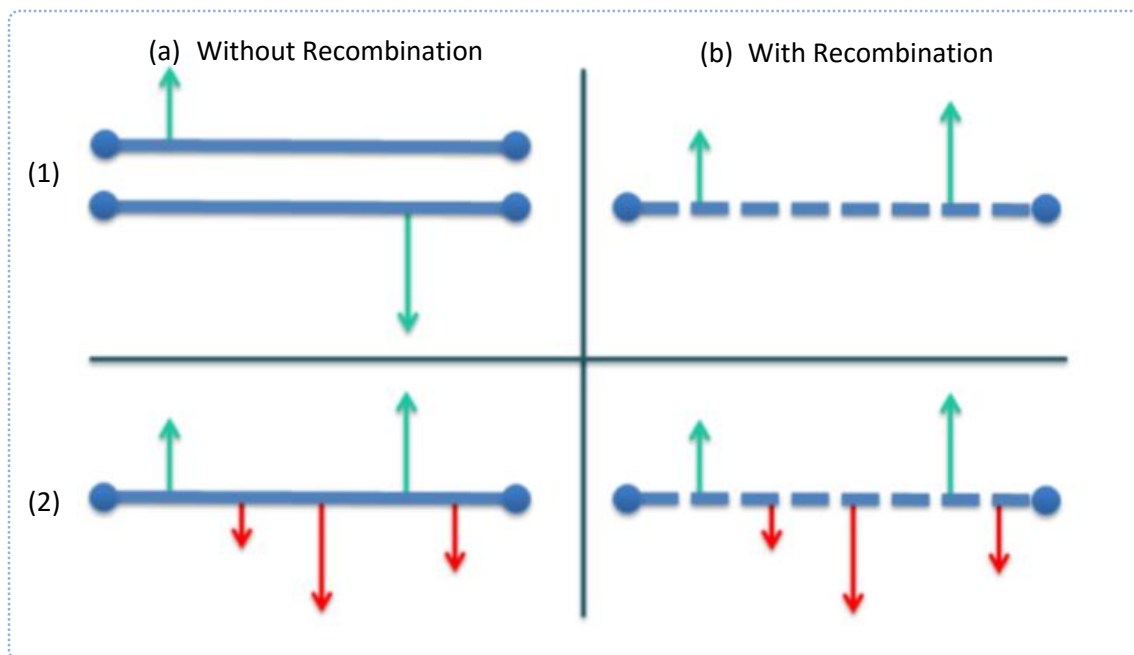
**Figure 1.5** Observed and predicted levels of polymorphism on the chromosome 3 of *D. melanogaster*. The observed data, from left to right, are from the following loci: Lsp1- $\gamma$ , Hsp26, Sod, Est6, fz, tra, Pc, Antp, Gld, MtnA, Hsp70A, ry, Ubx, Rh3, E(spl), Tl and Mlc2. The predicted  $\pi$  values are based on equation (8) of Hudson and Kaplan (1995) assuming that  $4N\mu = 0.014$  and  $u$ , the deleterious mutation rate per cytological band, is 0.0002. [Adapted from Barton et al 2007. Originally from Hudson and Kaplan 1995]

On the other hand, the effect of linked selection mechanisms in regions with low or no recombination makes selection inefficient and the mentioned mechanisms interfere between each other, as a consequence, variation is reduced in those segments of the genome. This situation, produced when various linked sites are selected simultaneously, has been called the Hill-Robertson interference (Hill & Robertson 1966, Begun and Aquadro 1992)(Figure 1.6).

There are two possible situations if various sites are mutually selected in a low recombination region: (i) two or more adaptive mutations appear in different haplotypes, both will compete



and only one will be fixed in the population, reducing the adaptive fixation rate (Figure 1.6 a1). Or (ii) there can be both adaptive and deleterious mutations in the same haplotype. The lack of recombination sometimes will lead to fixation of deleterious variants due to the fixation of a linked strong adaptive variant, or sometimes the opposite, where adaptive mutations will be eliminated if they are nearby a selected deleterious position (Figure 1.6 a2). The lower the recombination, the more sites that will segregate linked together. Moreover, the more intensity of selection, the more reduction of the efficiency of selection by the Hill-Robertson interference (Comeron *et al.* 2008; Messer & Petrov 2013). The interference does not happen if there is enough recombination that allows different nearby sites to segregate independently (Figure 1.6 b).



**Figure 1.6** Hill-Robertson interference in selected sites in a genome region. Arrows indicate selection on adaptive mutations (green) or deleterious (red). Length of the arrow indicates strength of selection. (1) Two or more adaptive mutations. Without recombination, both chromosomes compete and only one of the mutations become fixed (1a). If there is enough recombination, both mutations can be fixed (1b). (2) With presence of both adaptive and deleterious mutations, without recombination all alleles compete allowing deleterious fixations if selection on adaptive alleles is strong enough, or even adaptive mutations can be removed from the population if the selection in the deleterious sites is superior (2a). With recombination, deleterious alleles can be removed and adaptive alleles can be fixed together (2b). [From Barrón 2015]

### 1.1.5 The estimation of DNA variation

The *data desideratum* for population genetics studies is a set of homologous and independent sequences (or haplotypes) sampled in a DNA region of interest, along with the corresponding sequences from one or more outgroups to study both polymorphism and divergence. From a set of haplotypic sequences nucleotide diversity can be estimated for (i) each nucleotide site independently of other nucleotide sites (one-dimensional measure) or (ii) a segment of sites can be analyzed together taking into account the mutual associations among polymorphic sites (multi-dimensional measure) (Table 1.2). Nearby nucleotides are not independent from each other, since they tend to be clustered in blocks of different lengths, for example, up to 2kb in *Drosophila* (Miyashita and Langley 1988; Mackay *et al.* 2012) and over several megabases in the human genome (Frazer *et al.* 2007). Multi-dimensional estimators are important to describe the forces that shape haplotypes such as recombination, selection and demography. Both, one and multi-dimensional diversity measures, are complementary to get a complete description of sequence variation.

**Table 1.2** Common measures of nucleotide diversity

Uni-dimensional measures		
S, s	Number of segregating sites (per DNA sequence or per site, respectively).	Nei (1987)
H, $\eta$	Minimum number of mutations (per DNA sequence or per site, respectively)	Tajima (1996)
k	Average number of nucleotide differences (per DNA sequence) between any two sequences	Tajima (1983)
$\pi$	Nucleotide diversity: average number of nucleotide differences per site between any two sequences.	Nei (1987); Jukes and Cantor (1969); Nei and Gojobori (1986)
$\theta, \theta_w$	Nucleotide polymorphism: proportion of nucleotide sites that are expected to be polymorphic in any suitable sample	Watterson (1975); Tajima (1993; 1996)
Multi-dimensional measures		
D	The first and most common measure of linkage disequilibrium, dependent of allele frequencies	Lewontin and Kojima (1960)
D'	Another measure of association, independent of allele frequencies	Lewontin (1964)
R, R <sup>2</sup>	Statistical correlation between two sites	Hill and Robertson (1968)
ZnS	Average of R <sup>2</sup> over all pairwise comparisons	Kelly (1997)

[from Casillas 2007]

### 1.1.6 Detecting natural selection in the genome

One of the most amazing evidence of the power of natural selection is the footprint that it can leave on genetic variation. Looking for evidence of selection is also a widely-used strategy for finding functional variants in the genome (Bamshad and Wooding 2003). Several types of signatures leave natural selection in the genome: (i) a reduction in polymorphism, (ii) a skew towards rare derived alleles, and (iii) an increase in linkage disequilibrium (LD) (Bamshad and Wooding 2003). Several tests based on the level of variability and the distribution of alleles have been developed to identify the footprints of selection searching for such signatures (Table 1.3). However, it should be noted that several processes can interfere in the interpretation of these footprints.

*Hitchhiking events* reduce local levels of variation. Over time, since common neutral variants will have disappeared, new appearing mutations in the population start segregating at low frequencies leading to an excess of new rare derived alleles in the region. Also, a long region with high LD and low diversity can indicate recent positive selection over an allele if it is present at high frequency, since recombination still has not had enough time to reduce the LD (Figure 1.4).

*Background selection.* In a hitchhiking process the selected allele expands into the population along with other variants within its linked region. By contrast, in a background selection situation, different chromosomes carrying deleterious mutations are removed from the population, but no specific remaining variant is favoured. This leaves a more heterogeneous frequency spectra landscape with prevalence of rare alleles after a background selection event compared with the more homogenizing effect of a selective sweep (Figure 1.4).

Variation in the local rate of recombination along the genome also makes the detection of selection difficult, since the signatures of selection highly depend on the local rate of recombination (Hudson and Kaplan 1995). In this regard, the effects of non-selective processes like demography and recombination should be taken into account when trying to identify regions showing true signatures of evolution.

### 1.1.7 Tests of selection

**Table 1.3** Commonly used tests of neutralism

Test	Compares	References
<i>Based on allelic distribution and / or level of polymorphism:</i>		
Tajima's $D$	The number of nucleotide polymorphisms with the mean pairwise difference between sequences	Tajima (1989)
Fu and Li's $D$ , $D^*$	The number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants	Fu and Li (1993)
Fu and Li's $F$ , $F^*$	The number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences	Fu and Li (1993)
Fay and Wu's $H$	The number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies	Fay and Wu (2000)
SweepFinder	Detection of selective sweeps using composite likelihood	Nielsen <i>et al.</i> (2005)
<i>Based on comparisons of divergence:</i>		
$d_n/d_s$ , $K_a/K_s$	The ratios of nonsynonymous and synonymous nucleotide substitutions in protein coding regions	Li <i>et al.</i> (1985); Nei and Gojobori (1986)
PAML	Software suite that combines $d_n/d_s$ , phylogenetic, ML and Bayesian methods	Yang (2007)
<i>Based on comparisons of divergence and polymorphism between different functional sites :</i>		
HKA	The degree of polymorphism within and between species at two or more loci	Hudson <i>et al.</i> (1987)
MK	The ratios of synonymous and nonsynonymous nucleotide substitutions within and between species	McDonald and Kreitman (1991)
<i>Based on allelic distribution and comparisons of divergence and polymorphism:</i>		
DFE-alpha	Extended MK test using Site Frequency Spectrums to estimate the unbiased proportion of adaptive substitutions and distribution of fitness effects.	Keightley and Eyre-Walker (2009)
<i>Based on Linkage Disequilibrium:</i>		
EHH	Measurement of the decay of the association between alleles at various distances from a locus	Sabeti <i>et al.</i> (2002)
LHR	Test to search alleles of high frequency with long-range linkage disequilibrium	Sabeti <i>et al.</i> (2002)
iHS	Test to search for alleles under positive selection between shared haplotypes	Voight <i>et al.</i> (2006)
<i>Based on population comparisons:</i>		
$F_{st}$	Variance of allele frequencies between populations	Lewontin and Krakauer (1973); Akey <i>et al.</i> (2002)
XP-EHH	Extended haplotype homozygosity between populations	Sabeti <i>et al.</i> (2007)
XP-CLR	Search for quick changes in allele frequency in a region	Chen <i>et al.</i> (2010)

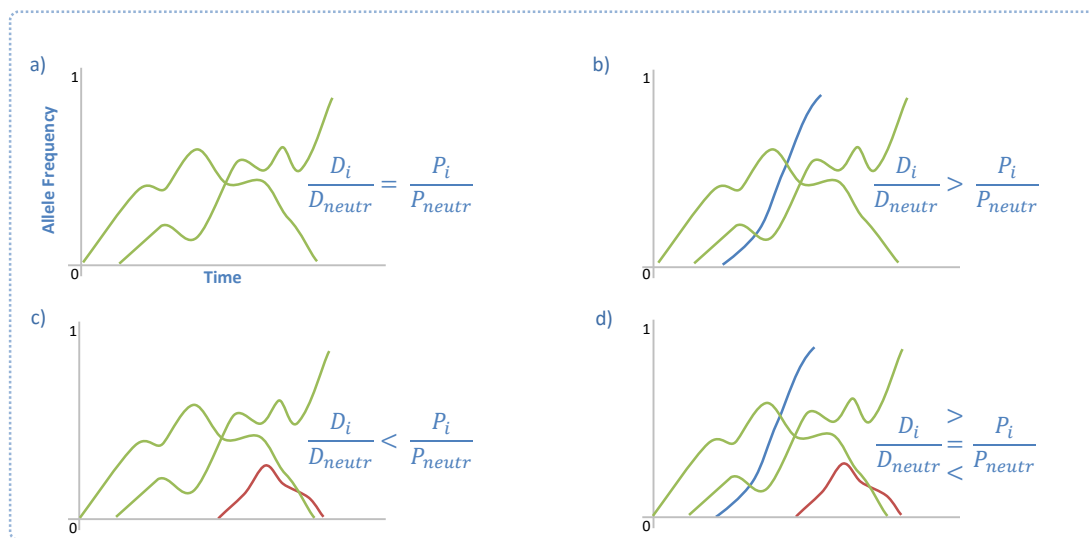
PAML, Phylogenetic Analysis by Maximum Likelihood; HKA, Hudson-Kreitman-Aguade; MK, McDonald-Kreitman; DFE, Distribution of fitness effects; EHH, Extended Haplotype Homozygosity; LHR, Long Haplotype Range; iHS, Integrated Haplotype Score; XP, Cross Population; CLR, Composite Likelihood Ratio .

In Table 1.3 are listed the commonly used tests for neutral pattern of variation in DNA data. They are classified according the kind data obtained: divergence data by comparing sequences of different species, polymorphic data from within population sequences, and data both from polymorphism and divergence.

**Tests based on levels of polymorphism.** One way to test if evolution is acting in a genomic region is to look at the polymorphism levels of different types of nucleotide sites and compare them with the expected levels of polymorphism in a neutral scenario. The ***K<sub>a</sub>/K<sub>s</sub> ratio test*** has been broadly used because the initial abundance of sequence data for different species,  $d_n/d_s$  (or  $K_a/K_s$ ) test (Yang and Bielawski 2000). In this test, the rate of nonsynonymous substitutions ( $d_n$  or  $K_a$ ) is compared to the rate of synonymous substitutions ( $d_s$  or  $K_s$ ) using the ratio  $\omega = d_n/d_s$ . The test assumes that (i) all synonymous substitutions are neutral and (ii) all substitutions have the same biological effect, which is not always true. If  $\omega > 1$  (for example, in a gene where a numerous nonsynonymous fixed mutations due to adaptive evolution have occurred) it's considered a signal for positive selection. On the contrary,  $\omega < 1$  (for example, if nonsynonymous mutations are being removed from a gene for being highly deleterious) is a signal of functional constraint. A powerful and exhaustive approach of the  $d_n/d_s$  method is found in the software package PAML 4 (Yang 2007), which combines phylogenetic, maximum likelihood (ML) and Bayesian methods.

**McDonald-Kreitman test (MKT).** The MKT (McDonald and Kreitman 1991) compares divergence ( $D$ ) between species and polymorphism ( $P$ ) inside a species at two types of sites. At least, one site class must be a putatively neutral class ( $P_s, D_s$ ) which is compared with the other site class to test if it's under selection or not ( $P_i, D_i$ ). Designed initially for coding sequence analysis, synonymous sites were the classical putatively neutral class and non-synonymous positions the ones to test if they were under selection or not. If all the mutations are either neutral or strongly deleterious, then  $D_i/D_s$  is expected to be very similar to  $P_i/P_s$ . On the other hand, a case of positive selection would imply more fixations and it would be reflected as more divergence compared to the polymorphism ( $D_i/D_s > P_i/P_s$ ). On the contrary, an excess of polymorphism with respect to the divergence ( $D_i/D_s < P_i/P_s$ ) would be signal of deleterious alleles segregating in the population, which are lost preferentially and therefore underrepresented as divergence substitution. The MKT can potentially be generalized to test 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 (Egea *et al.* 2008).

Two features make the MKT especially useful to infer selection: (1) the use of polymorphism and divergence data can avoid the confounding effects of other evolutionary processes such as mutation or recombination rate. The inequality of both ratios ( $P_i/P_0 \neq D_i/D_0$ ) cannot be attributed to mutation rate differences between both sites, because it will affect both ratios equally. Likewise, the MKT allows separating mutation associated with recombination rate from selection as causes of excess of variation in highly recombinant regions; (2) granted that the two classes of sites are closely linked, they share a common evolutionary history, which makes the MKT remarkably robust to assumptions about non-equilibrium demography (Nielsen 2001; Eyre-Walker 2002) and recombination rates (Sawyer and Hartl 1992).



**Figure 1.7** Different scenarios that can be discovered by the MKT. (a) If only neutral alleles (green) exist in the population, we expect an equal proportion of divergent and polymorphic sites as MKT result. (b) We expect an excess of divergence compared with the polymorphism due to faster fixation of adaptive alleles (blue). (c) On the contrary, if there are slightly deleterious alleles (red) in the population we observe an excess in polymorphism since these alleles can segregate for a time before being removed. (d) However, if both slightly deleterious and adaptive alleles are present, the results of the MKT can be easily misinterpreted.

Assuming that adaptive mutations seldom contribute to polymorphism and are detected only as divergence, the proportion of adaptive substitutions ( $\alpha$ ) can be estimated ( $\alpha = 1 - [D_s P_i / (D_i P_s)]$ ) (Charlesworth 1994). However, one main concern of the MKT refers to the presence of deleterious segregating alleles. Since the test assumes that all non-synonymous mutations are either strongly deleterious, neutral or strongly advantageous, estimates can be easily biased by the segregation of slightly deleterious nonsynonymous mutations, and adaptive selection can severely be underestimated (Eyre-Walker 2002)(Figure 1.7). The exclusion of low frequency polymorphisms (Fay *et al.* 2001) has been used to detect adaptive selection as it increases the

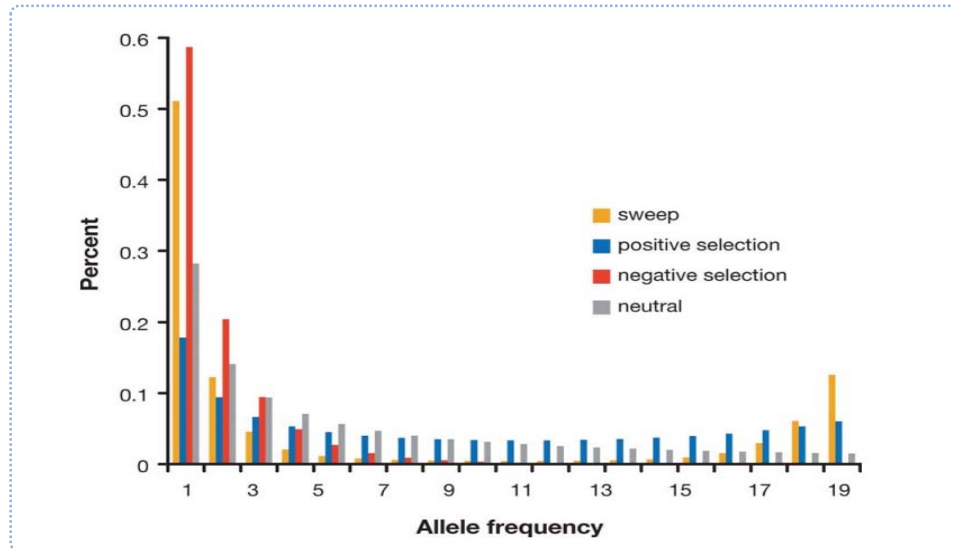
power of the MKT (Andolfatto 2005, Charlesworth and Eyre-Walker 2008), however this may also make the test more sensitive to artifactual evidence of adaptive evolution if there has been an increase in effective population size, since some slightly deleterious mutations not segregating in the population may have been fixed in the past (Eyre-Walker 2002).

**Tests based of the site frequency spectrum (SFS).** The use of SFS, which assigns the number of segregating alleles within a given frequency class, is another approach to test the neutrality of a genomic region. Different evolutionary events can leave a distinctive pattern in the SFS (Nielsen 2005). Neutral variation offers a spectrum with a fairly high number of alleles in lower frequencies, which continuously decreases as the frequency increases. Slightly deleterious mutations presents an excess of low frequency variants when compared with the neutral spectrum. On the other hand, events of positive selection are detected with a reduction of low frequency variants and increased number of variants at middle and high frequencies. SFSs are also useful to detect sweeps since they usually leave a pattern of excess of both low and high frequency variants with a severe reduction of middle frequency variants (Figure 1.8) (Nielsen 2005).

**Fay and Wu's  $H$  and Fu and Li's  $D$  and  $F$  tests** are traditional neutrality tests that use SFS, but many other tests have appeared after them (Achaz 2009). **SweepFinder** (Nielsen *et al.* 2005) is a test based on calculating a composite likelihood to detect selective sweeps using SFSs inferred from SNP data, which excludes biases due to demographic effects or changes in mutation or recombination rates.

Another recent test relevant for this thesis is the **DFE-alpha** (Keightley and Eyre-Walker 2009). It actually uses the SFSs and neutral expected versus observed comparisons after Monte Carlo simulations to extend the MKT and try to correct for the slightly deleterious and demographic biases. DFE-alpha models the DFEs for putatively neutral and selected class sites by mean of a gamma distribution, which depends on two parameters: (i) the mean strength of selection ( $\gamma$ ) and (ii) a shape parameter ( $\beta$ ). The method simulates two demographic situations: (i) constant population size and (ii) a single, instantaneous change in population size from an ancestral size ( $N_1$ ) to a present day size ( $N_2$ ) that occurred ( $t$ ) generations ago and infer the adaptive substitution rate ( $\alpha$ ) for the putatively selected class.

Mackay *et al.* (2012) have also developed a modification of the MKT using SFS data, the **Integrative MKT**, which allows the estimation of five different regimes of selection from polymorphic and divergence data (see Box 2).



**Figure 1.8** Example of frequency spectra under a selective sweep, negative selection, neutrality and positive selection [from Nielsen 2005]

### Box 2: The integrative McDonald-Kreitman test

The integrative-MKT is a method that incorporates site frequency spectrum data to the framework of the MKT to taken into account weakly deleterious alleles segregating in the population (Mackay *et al.* 2012, supplementary materials). The Integrative-MKT allows estimating five different regimes of selection acting on new mutations (Table 1.4) for any given region of the genome. Adaptive mutations and weakly deleterious selection act in opposite directions on the MKT, so if both selection events are occurring, they will mutually be underestimated (Figure 1.7). To take both adaptive and slightly deleterious mutation into account,  $P_i$  (the count of segregating sites in the selected class  $i$  of the standard MKT table) has to be decomposed into the number of neutral variants and the number of weakly deleterious variants ( $P_i = P_{i\text{neutral}} + P_{i\text{weakly del.}}$ ). From the SFS of neutral sites,  $P_{i\text{neutral}}$  can then be estimated and five regimes of selection (Table 1.4). The integrative-MKT has been implemented in software written in Java.

**Table 1.4** Estimated regimes of selection by the integrative-MKT

Selective fraction	Symbol	Estimator
Strongly deleterious sites	$d$	$\hat{d} = 1 - (\hat{f} + \hat{b}) = 1 - (m_0 p_i / m_i p_0)$
Weakly deleterious sites	$b$	$\hat{b} = (\hat{P}_{i\text{weak del}} / P_0) (m_0 / m_i)$
Neutral sites	$f$	$\hat{f} = (m_0 \hat{P}_{i\text{neutral}} / m_i P_0)$
Sites that have become neutral (subset of $f$ )	$\gamma$	$\hat{\gamma} = [(\hat{P}_{i\text{neutral}} / P_0) - (D_i / D_0)] (m_0 / m_i)$
Adaptive fixations	$\alpha$	$\hat{\alpha} = 1 - (\hat{P}_{i\text{neutral}} / P_0) (D_0 / D_i)$



**Tests based on linkage disequilibrium.** A typical signal of a sweep event caused by positive selection is a long region with high linkage disequilibrium and reduced polymorphism. Methods like the **Extended Haplotype Homozygosity (EHH)**, the **Long Range Haplotype (LRH)** or the **Integrated Haplotype Score (iHS)** try to detect and assess such signals (Sabeti *et al.* 2002, Voight *et al.* 2006). Finally, some methods incorporate the **Fixation Index ( $F_{st}$ )**, a statistic that describes the differentiation between populations using allele frequencies (Lewontin and Krakauer 1973, Akei *et al.* 2002). The **XP-EHH** (Cross Population Extended Haplotype Homozygosity) (Sabeti *et al.* 2007) combines the previously mentioned EHH test with comparisons between populations to search for alleles under positive selection. In a similar way, the **XP-CLR test** (Cross Population Composite Likelihood Ratio) (Chen *et al.* 2010) try to search for alleles under positive selection not by looking at LD levels but with changes in allele frequency.

#### 1.1.8 Detecting selection genome-wide

Until recent years, population genetics studies have been so far based on fragmentary and non-random samples of genomes, providing a partial and often biased view of the population genetics processes (Begun *et al.* 2007). In the case of selection tests, traditionally, most of them only compared specific sets of variants against neutral empirical or simulated expectations.

In recent years, the abundance of genomic data due to the high throughput of NGS and the rise in computational power allowed to test not only regions, but to scan complete genomes for selection signatures (Oleksyk *et al.* 2009). Genome-wide scans for selection usually use either re-sequencing data from one or many species (Bustamante *et al.* 2005) or large collections of SNP data like the HapMap in humans (Altshuler *et al.* 2005; Frazer *et al.* 2007). This availability of data and the computational capacity to analyze it massively has allowed applying the methods to detect selection explained in the previous section not only to particular regions but complete genomes. This change of scale describing variation has made population genetics to become population genomics.

## 1.2 Population genomics: population genetics meets genomics

---

A global-genome view of diversity allows re-addressing questions in population genetics whose response was uncertain in previous studies because potential sources of bias are uncontrolled when sampling specific genes or regions of the genome. Consider the correlation found between the level of polymorphism and divergence at any given region. Because this correlation may vary according to the chromosome region considered, any average estimate that does not track the patterns along the whole chromosome arms could be biased. A global perspective lets us detect differences in variation patterns among and within chromosome arms, as for example between autosomes and sexual chromosomes. A genome-wide analysis allows then monitoring local versus regional effects along chromosome arms to decipher the role of recombination rate, mutation rate or gene density on the amount of nucleotide variation and/or adaptive evolution.

Essential for population genomics studies has been the model organism *Drosophila melanogaster*. The fruit fly *D. melanogaster* is one of the most successful experimental model used in the laboratory (Roberts 2006). With a genome size of ~176Mb, ~5% the size of a mammal genome on average, *D. melanogaster* still shares with mammals many gene families, pathways and tissues (De Velasco *et al.* 2004, Kida *et al.* 2004). Since first used by Morgan during the first years of genetics (Morgan *et al.* 1915), it has assisted research in many fields of biology, especially in genetics and development biology. Its relevance even made *D. melanogaster* to be the third eukaryotic genome ever sequenced, after the yeast *Saccharomyces cerevisiae* (Goffeau *et al.* 1996) and the nematode *Caenorhabditis elegans* (Consortium 1998). Moreover, the fruit fly was selected to be the first eukaryotic organism to test whole genome shotgun sequencing (WGS) (Rubin 1996, Adams *et al.* 2000), a crucial step that afterwards led to the present NGS methods.

The study of Begun *et al.* (2007) in *Drosophila simulans* can be considered the first true population genomic dataset (Hahn 2008) closely being followed by the Liti *et al.* (2009) population genomics study in wild and domestic yeast. However, the mentioned study in *Drosophila simulans* (Begun *et al.* 2007) and a following in *D. melanogaster* (Sackton *et al.* 2009) were based on low-coverage sequencing. In the *D. simulans* project, from the seven lines analysed, only 6 were considered due to a mixing of samples, and from those, the average coverage was on average 3.9. In the *D. melanogaster* study the mean lines aligned only rises up to 5.4. These are values manifestly insufficient for any population genetic inference based on

frequency of variants where it's suggested a minimum coverage of 10-15x using short reads technologies (Craig *et al.* 2008, Smith *et al.* 2008).

The results of Begun *et al.* (2007) work have challenged the traditionally and widely accepted explanation of neutral theory (Hahn 2008). These works, despite their sample limitations, opened the path to the next big population genomics studies. One of these studies was carried out on the *Drosophila Genetic Reference Panel* (DGRP), whose data has been used to develop the present thesis project.

### 1.2.1 The *Drosophila* Genetic Reference Panel

The DGRP is an international effort with the objective of the complete characterization, both in genotype and phenotype, of around two hundred lines sampled from a natural population of *Drosophila melanogaster* in Raleigh, North Carolina (USA) (Mackay, Richards and Gibbs 2008). The main goals of the DGRP are the creation of: (i) a community resource for association mapping of quantitative trait loci (QTL) for traits relevant to human health. (ii) A community resource of common *Drosophila* sequence polymorphisms for its use in QTL mapping and population genomics analysis. (iii) A test bench for statistical methods used in QTL association and mapping studies.

It's known that *D. melanogaster* is a recent cosmopolitan species, whose origin can be traced in Africa (Lachaise *et al.* 1988, David and Capy 1988, Begun and Aquadro 1993, Andolfatto 2001, Stephan and Li 2007, Duchon *et al.* 2013). This also makes *D. melanogaster* interesting to study the evolutionary implications of large migrations, especially with the availability several other *Drosophila* genomes (Consortium 2007) from around the world to compare different evolutionary histories, and also for the parallelism with the human species. In this regard, the Raleigh population is especially interesting, since it seems that *D. melanogaster* arrived in America less than 200 years ago (Lintner 1882, Keller 2007).

One problem that arises when trying to genotype diploid species like *D. melanogaster* is to distinguish real heterozygous sites at the same locus from distinct paralogs loci (Vinson *et al.* 2005). Also, the presence of heterozygous sites makes difficult the distinction between real polymorphism and sequencing errors. Three strategies are followed to deal with this problem: (i) the creation of inbred pure lines to increase the proportion of homozygous sites; (ii)

sequencing haploid embryos, obtained from the offspring of a female mated with a male homozygous for a deleterious allele in the locus *ms(3)K1* which causes mitotic failure of the paternal chromosomes during the first rounds of cell division; (iii) the use of balancing chromosomes and chromosome extraction (Langley *et al.* 2011).

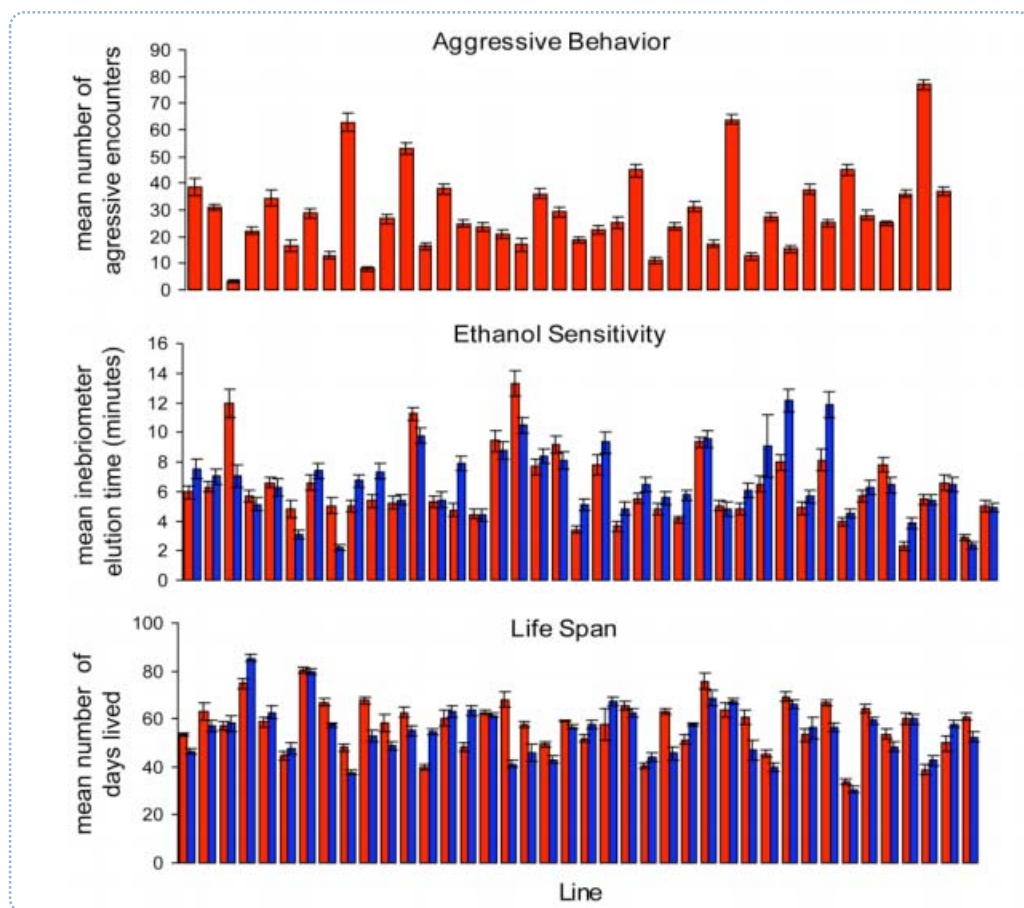
The DGRP data gathering and analyses had contemplated three phases so far:

1. Initial phase: A white paper was presented in which 40 lines were characterized phenotypically and genotypically to test the viability of the project (Mackay, Richards and Gibbs 2010).
2. Freeze 1.0 (February 2012): Sequencing of 168 inbred lines. 129 lines were sequenced with Illumina technology, 10 lines with 454 technology and 29 with both 454 and Illumina. Illumina reads had an average 21X coverage per line while for 454 reads had 12.1X coverage per line. Only SNPs were genotyped and used for QTL and population genomics analysis (Mackay *et al.* 2012).
3. Freeze 2.0 (July 2014): Sequencing of 205 lines (including Freeze 1.0 lines) with longer read Illumina technologies, with coverage of 27X per line on average. SNP and non-SNP variation was genotyped and used for QTL and population genomics analysis (Huang *et al.* 2014).

In this work, the 158 genomes of *D. melanogaster* together with the genome sequences of its closest species, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*, have been described and analyzed by means of a battery of comparative methods for polymorphism and divergence data to answer questions of fundamental interest in population genomics such as: Which pattern or gradient follows genetic variation along the chromosomes? How these patterns correlate with structural regions? Which proportion of the coding and non-coding genome undergoes purifying, neutral or positive selection? How recombination rate determine nucleotide variation and molecular evolution along the genome?

Finally, even though it has not been part of the research done in this thesis, one main goal of the DGRP is to decipher genotype-phenotype relationships and try to create the most fine scale genotype-phenotype study to date. Interactions between genotype and phenotype are complex, and still poorly understood, but the huge amounts genomic data and computational power can help us to shed light about the processes governing these interactions (Figure 1.9).

Trying to understand and to define the genotype-phenotype map is the core aim of the DGRP, and this is of paramount importance to understand the causal path of natural selection, since it acts primarily on the phenotype and only indirectly, as a function of the genotype-phenotype map, on the genotype (Lewontin, 1974). This is something that molecular population geneticists tend to forget in this era of fascination with genome data, that what it's "ultimately to be explained are the myriad and subtle changes of size, shape, behavior, and interactions with other species that constitute the real stuff of evolution" (Lewontin 1974). This certainly are big steps towards a more integrated way of study complex traits in what has been recently called the *Systems Genetics* approach (Civelek and Luskis, 2014).



**Figure 1.9** Variation in three quantitative traits in 40 of the proposed lines of the *Drosophila* genetic reference panel. Red: males; Blue: females. [From Mackay, Richards and Gibbs (2008)]

## 1.3 Genome Browsers

---

### 1.3.1 Molecular Databases

The explosion of genome sequence data in the last decade has been so widely cited as to have almost become a *cliché* (Schattner 2008). The first microbial genome was sequenced just in 1995. Similarly, the first complete genome of a multicellular organism (*C. elegans*) became available in 1998. The rate at which genomes for new species and within species individuals are being sequenced continues to accelerate as novel sequencing technologies lower the cost of obtaining sequence data. This is clearly observable in web sites like Genomes Online Database (GOLD) that tracks genome sequencing projects, and that at the current date (April 2015) counts 58693 sequenced genomes of organisms, from which 1,037 corresponds to archaea domain, 44,576 are eubacteria and 8,181 are eukaryotic genomes.

A helpful way to use this data for biological research has been organizing it into dedicated databases. However, as the number of databases keeps growing, integrating and extracting knowledge from them becomes really challenging. The biological research community has also brought even more difficulties into this task, especially by the way this data has historically been stored: many databases that are only downloadable as flat files, relational databases that need to be set up locally or varying data formats that need different parsers and convertors. All those factors make comparing and integrating different biological data sources difficult and tedious.

Genome databases offer solutions to these problems. By aggregating data from multiple databases and integrating data in a uniform and standardized manner, they enable researchers to formulate complex biological queries involving data that were originally from diverse sources (Schattner 2008). By a genome database we mean a data repository, generally implemented via relational databases, that include the maximum available genomic sequence data of one or more organisms, together with additional information that are usually referred as *annotations*.

The creation of a genomic database is a complex, and usually a multitask endeavour that can be summarized in these fundamental tasks (Schattner 2008):

- Sequencing the genomic DNA
- Assembling the fragments of DNA sequence data into continuous pieces spanning all or most of the length of the organism's chromosomes
- Aligning transcript data to the genomic sequence

- Identifying the locations of the genes within the genome sequence
- Designing and implementing the data-storage architecture to house the data
- Maintaining and updating the database as additional data become available

Once a genome is successfully sequenced the next step is to identify and describe any functional region. The process to add biological information into a sequence is called annotation (Stein 2001). Identifying functional regions can be done in the lab, a process called *manually curation*, or automatically using bioinformatic prediction tools. Over the years, many software has been developed to predict genes and other functional regions using different approaches: from complex pattern searches into the sequence, to the integration of NGS read information like RNAseq to detect regions being actively transcribed. In general, manually curated annotations yield fewer false positives than purely computational approaches, but are more labour intensive and tend to generate more false negatives than automated methods. Genomic databases can contain either one of these types of annotations or both.

Annotation databases are diverse: for functional sequences, proteins, pathways, short reads, etc. However, the trend is to try to integrate the major number of databases into general portals to aid the search work of the researchers. 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>). Most of the annotation types that can be found in a genomic database are summarized in Table 1.5.

**Table 1.5** Common annotation types found in genomic databases

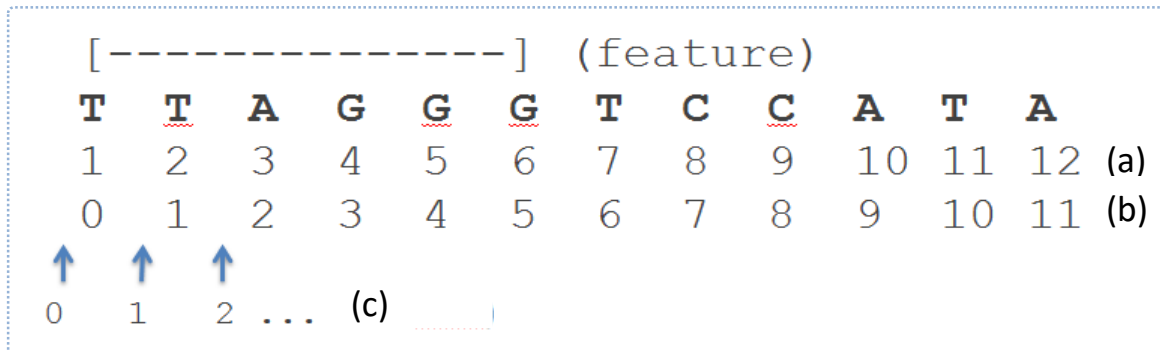
Associated with a genomic coordinates	Not associated with a genomic coordinates
<ul style="list-style-type: none"> <li>• Locations of genes</li> <li>• Gene-structure annotations indicating a gene's exon-intron boundaries</li> <li>• Locations of known and putative gene regulatory regions such as promoters, transcriptional enhancers, CpG islands, splicing enhancers and silencers, DNase hypersensitive sites, nucleosome sites, and so on</li> <li>• Transcript alignments indicating the genomic sources of observed proteins, mRNAs/cDNAs, and expressed sequence tags (ESTs)</li> <li>• Alignments of protein, mRNA, and EST sequences from related species</li> <li>• General chromosomal features such as repetitive sequences, recombination "hotspots," and variations in local CG%</li> <li>• Alignments of genomic DNA from other species, which can provide clues regarding sequence conservation and chromosomal evolution</li> <li>• Annotations of regions that vary within a population of individuals, including single nucleotide polymorphisms (SNPs), short indels, large structural or copy number variations, and correlations among sequence variations, such as those that have been identified by the haplotype mapping projects (e.g., HapMap)</li> <li>• Genome-wide RNA expression data from multiple sources</li> <li>• Sequence features that are used in the process of assembling the genome, such as sequence tagged sites (STSs) from genetic and radiation hybrid maps, NGS normal or paired-end reads.</li> </ul>	<ul style="list-style-type: none"> <li>• Protein structure data</li> <li>• Evolutionary data, including evolutionary relationships among individual genes as well as among chromosomal regions and entire genomes</li> <li>• Annotations describing phenotype variations</li> <li>• Metabolic- and signaling-pathway data</li> <li>• Protein-interaction data, such as data from yeast two-hybrid system experiments</li> <li>• and data derived from protein-chip expression analysis</li> </ul>



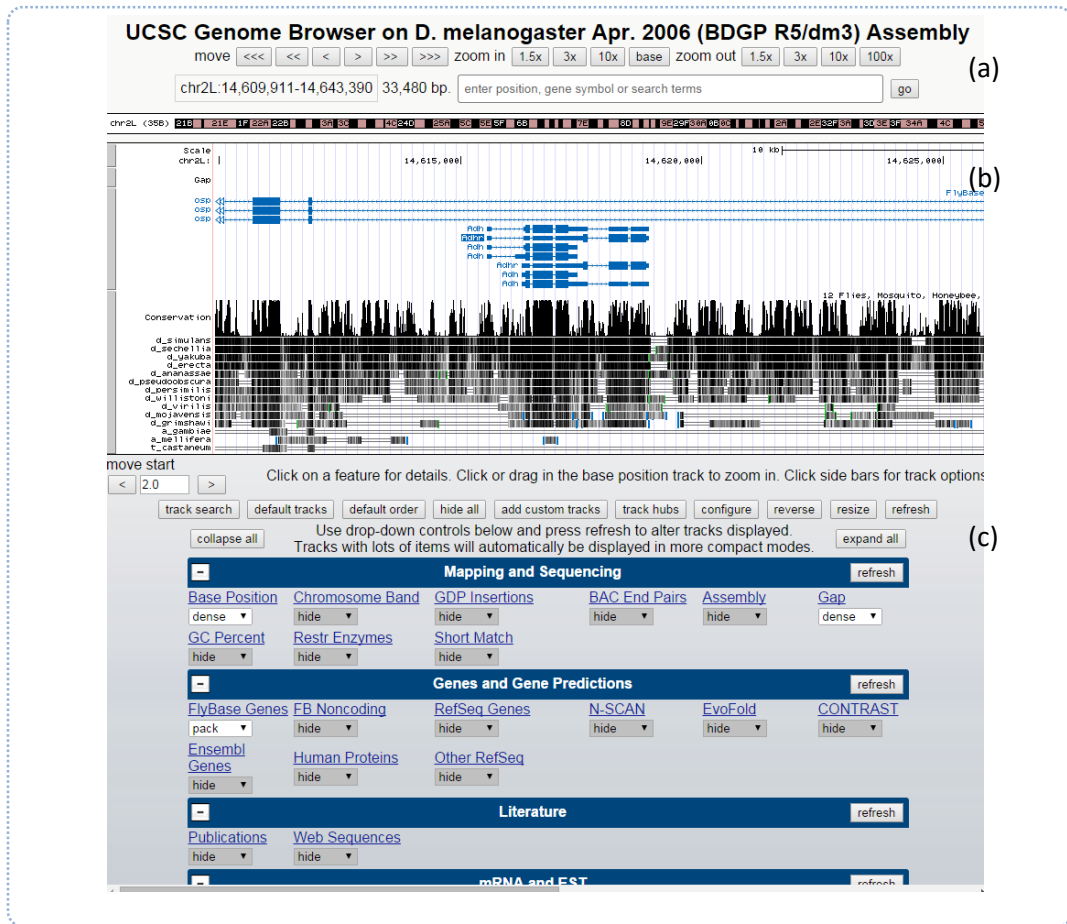
### 1.3.2 Genome Browsers

As the number of genomes and annotations grow, it does also the need for a dynamic, flexible system to store, modify and retrieve all this information. Biological databases already are part of the everyday tools used by many biologists, even the ones not dealing with bioinformatics work. The visual access to genomic information via genome browsers was one of the many ways that genome database creators implemented in their solutions.

We can define a Genome Browser as a tool to visually access a given annotation database. And they offer a flexible way to rapidly visualize annotations, not independently, but in their own genomic context. Generally, a reference genome is used as a coordinate system (Figure 1.10) where annotations are anchored. Basically a genome could be understood as one dimensional map and the annotations are the landmarks. So, any genome browsing system should provide navigation tools to move back and forth, or zoom in and out through arbitrary regions of a genome. The way to visualize annotations in a region is using *tracks*, non-overlapping layers of information of the corresponding region, where graphical representations of the annotations or *glyphs* are displayed (Figure 1.11).



**Figure 1.10** Different annotation coordinate systems. Depending on the database used, annotation coordinates can differ if the first nucleotide in a genome is considered position 1 or position 0. This also has implications in the annotation of 1bp features (for example an SNP) or features without length (like an inserted segment that does not exist in the reference genome, thus, it has no coordinates of its own). In a 1-based system (a) an SNP in the third position is stored as 3 for the start coordinate and also 3 for the end coordinate. An insertion between the third and fourth position is stored as start = 3 and end = 2. In a 0-based coordinate system the same SNP has start = 3 and end = 4 coordinates while the insertion has start = 3 and end = 4 coordinates. There's a third coordinate system, the interbase system (c), that does not count nucleotides but the spaces between them, but in practice it works the same as a 1-based system since the space with coordinate 0 is the one previous to the first nucleotide. [Adapted from Schattner 2008]



**Figure 1.11** Snapshot of the UCSC Genome Browser in the region of the *Adh* gene in *D. melanogaster* with the main sections of a genome browser interface. We can see (a) the navigation control buttons, (b) a main section where the tracks and glyphs are displayed and (c) the track selection section.

However, there exist a lot of genome browser tools since the bioinformatics research community seems to reinvent the same applications again and again during these past years (Stein 2002). Current genome browsers can be classified in many ways (Table 1.6). One way is separate the ones that are deeply integrated with their data (data warehouses) from the ones that are initially 'empty' and it's the user who must add the annotations to display (generics). Generally, data warehouse browsers are not designed to be portable; they cannot work outside the hardware and software infrastructure of their database. Another classification is the distinction between desktop applications and web based genome browsers.

**Table 1.6** Most relevant non-proprietary genome browsers

Browser	Description	Database Type	Interface Type
MapViewer	Genome browser for the NCBI databases	Data Warehouse	Web Based
UCSC Genome Browser	Browser for the University of California Santa Cruz genome databases.	Data Warehouse	Web Based
ENSEMBL	European Bioinformatics Institute (EMBL) genome browser	Data Warehouse	Web Based
GBrowse	Genome browser developed by the GMOD (Generic model organism databases) community. Used, for example, in the FlyBase and HapMap portals.	Generic	Web Based
IGV Browser	Desktop genome browser developed by the Broad Institute.	Generic	Desktop
GenomeMaps	Genome browser developed to make use of modern web programming languages and HPC infrastructures.	Data Warehouse	Web Based

### 1.3.3 GMOD community and the Generic Genome Browser

In the mess of approaches to the genomic data analysis, any initiative with the goal to clarify and standardize procedures is more than welcome. The *Generic Models Organism Databases* (GMOD) Community is, maybe, the most remarkable bioinformatics initiative in this regard. GMOD is basically a collection of open source genomic data analysis tools and a network between its developers and users. The repository comprises well known tools by the community such as the Apollo and Maker (Lee *et al.* 2009, Holt and Yandell 2011) annotation tools, the Chado database framework (Mungall *et al.* 2007), the BioMart data mining toolset (Kasprzyk 2011), the Galaxy tool integration and workflow manager tool (Goecks *et al.* 2010) or the Generic Genome Browser (GBrowse) (Stein *et al.* 2002).

GBrowse is a web-based application for displaying genomic annotations and other features (Stein *et al.* 2002). From the administration side, GBrowse is mainly an integration of BioPerl modules (Stajich 2002) working under Unix systems with a running web server software. GBrowse can use common annotation formats such as GFF, BED, GFF or WIG to display features in a given genomic region through an HTML/JavaScript web interface. Annotations can be loaded into relational databases like MySQL or PostgreSQL directly or using the Chado framework. Recently, NGS data display support has been added using SAM/BAM files and the software SAMtools (Li 2009). As an open source tool developed in Perl, the administrator has the possibility to expand GBrowse functionalities with custom code via a plug-in system. The HTML

interface is highly customizable with CSS and Javascript custom code. For the end user, features of the browser include the ability to scroll and zoom through arbitrary regions of a genome, to enter a region of the genome by searching by landmark, the ability to enable or disable tracks a change its order and appearance, the possibility to upload custom data and data download and sharing capabilities.

As its name implies, GBrowse main feature is its *generic* nature. This means that the application is not packaged with any mandatory data to display and that the administrator has the freedom to create any genomic database. This has promoted GBrowse to be used in multiple genome database projects being the most remarkable the Human polymorphism HapMap browser (HapMap consortium 2003), The J. Watson's individual genome project (Wheeler 2008), The *Drosophila* portal FlyBase (St Pierre *et al.* 2014), the *Caenorhabditis* portal WormBase (Yook 2012), the Mouse Genome Informatics - MGI database (Blake 2014), The *Arabidopsis* Information Resource - TAIR (Lamesch 2011) or the *Saccharomyces* Genome Database - SGD (Cherry 2012).

## 1.4 Objectives

---

This thesis is both a population genomics study and a bioinformatics project centred on the visualization, description and analysis of the genome-wide DNA variation of a natural population of *Drosophila melanogaster*. The objectives of this project are (i) the description of the genome-wide nucleotide variation, (ii) the description of common non-SNP variation and (iii) the visual representation of such variation.

- I. **Population genome browser.** As part of our contribution in the DGRP project we aimed to create an online map of the genome polymorphism in the *Drosophila melanogaster* in open access to the scientific community. We use available open source tools that allowed us the addition of some new functions useful for genome wide population description, analysis and query of DNA variation.
  
- II. **Description and interpretation of Genome-wide SNP diversity.** First we will describe the nucleotide variation patterns across the chromosomes arms of *D. melanogaster* from the DGRP lines by using a sliding window approach. Then we will try to infer the population genetics processes responsible of the variation distributions, aiming to find and explain differences in variation between regions and chromosomes. Using this variation data set, standard and new methodologies to search for footprints of natural selection genome-wide will be applied, and the role selective and non-selective forces shaping the variation patterns in the *D. melanogaster* genome will be assessed.
  
- III. **Description and interpretation of Genome-wide non-SNP diversity.** Using the recently available Freeze 2.0 data of non-SNP variation in the DGRP population we aim to perform a variation analysis complementary to the one of SNP variation. We describe the genome-wide distribution of non-SNP variations in a similar way to the SNP variation of the previous objective. Moreover, we try to describe how SNP and non-SNP variation patterns are related in the genome and to infer the selective forces impinging on it.



---

**Part 2.**  
**Materials and methods**

---

## 2. Materials and Methods

### 2.1 DGRP Input Data

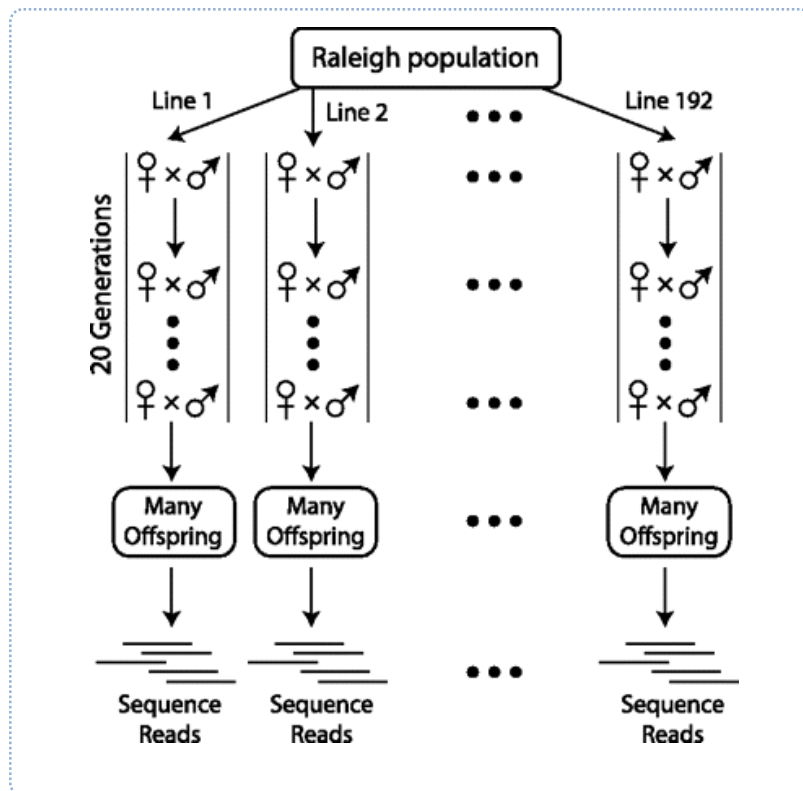
---

The initial data used in this thesis project are the sequences and variants called from the inbred lines of the DGRP (see 1.2.1). The DGRP project had one initial test phase and two working phases named ‘Freeze 1’ and ‘Freeze 2’.

The DGRP population was created collecting gravid females and following the full-sibling inbreeding approach during 20 generations to obtain full homozygous individuals. After this number of generations it is expected to have ~1.4% of residual heterozygosity in the samples (inbreeding coefficient  $F=0.986$ , Falconer and Mackay 1996). To call variants correctly, the ‘Joint Genotyper for Inbred Lines’ (**JGIL**, Stone 2012), a novel genotype caller that takes into account inbreeding, was developed specially for the DGRP. JGIL takes into account coverage, site-specific errors, quality sequencing statistics, and expected allele frequencies after 20 generations of inbreeding from an outbred population initially in Hardy–Weinberg equilibrium.

After genotyping, the expected ~1.4% of residual heterozygosity was true for ~90% of the lines. DGRP lines showing high values of residual heterozygosity (>9%) were observed to be associated to large polymorphic inversions. Heterozygous sites could be maintained due to a higher fitness for some heterozygous loci or to the presence of recessive lethal loci (Huang *et al.* 2014). Because  $2N_e = 4$  during the full-sibling inbreeding procedure, only lethal or strongly deleterious alleles are expected to be purged (García-Dorado *et al.* 2012), thus we expect that the inbred lines contain a rather representative random sample of the natural variation present in the population at the moment at which the flies were sampled.





**Figure 2.1** Experimental design to obtain and sequence the DGRP lines. Each line was founded by a gravid female collected from the Raleigh, North Carolina Farmer's Market (USA). Each subsequent generation was created by crossing a pair of male and female progeny from the previous generation. The DGRP lines were found after 20 generations of full-sib inbreeding. For each line, high-throughput sequencing was performed on DNA that was extracted from a pool of 500–1000 flies. (From Stone 2012).

### 2.1.1 Sequence data

**Freeze 1.** The initial input data is a set of 158 intraspecific *D. melanogaster* whole genome sequences provided by the DGRP project, in multi-Fasta file format. The alignments come from an initial sequencing of 168 DGRP lines using Illumina and 454 technologies (see 1.2.1). Illumina sequences had an average of 21X coverage per line and 454 reads had 12.1X coverage per line. Only Illumina lines sequences were used to reconstruct the sequences.

Illumina sequence reads were aligned to the *D. melanogaster* 5.13 reference genome using BWA (Li and Durbin 2010), duplicates were removed with GATK (McKenna *et al.* 2010). Finally, JGIL was used to validate SNPs for each line (Mackay *et al.* 2012, Stone 2012).

The four genomes phylogenetically closest to *D. melanogaster* which were sequenced by the Drosophila 12 genomes consortium were considered as outgroup species for the analyses of divergence: *D. simulans* (a mosaic of several different *D. simulans* lines), *D. sechellia* (4.9X coverage), *D. yakuba* (9.1X) and *D. erecta* (10.6X). Files in format axtNet containing the

alignment blocks of the genome of *D. melanogaster* with each one of the species were downloaded from the UCSC genome browser. The alignment blocks were arranged and merged to align *D. melanogaster* whole chromosomes using ad hoc scripts written in Perl. Finally, most of the analyses requiring an outgroup sequence were performed using *D. simulans* and *D. yakuba*.

**Freeze 2.** Along this phase, 205 DGRP lines (157 already sequenced in the previous freeze) were sequenced with longer read Illumina technologies and 27X coverage. Sequences were aligned with BWA and Novoalign (Novocraft.com), duplicates removed with GATK (Huang *et al.* 2014). A total of 4,853,802 unique SNPs and 1,296,080 non-SNP variants were called using multiple calling software: GATK, JGIL, Atlas-SNP (Shen *et al.* 2010), PrinSeS (Massouras *et al.* 2010), DELLY (Rausch *et al.* 2012), Pindel (Ye *et al.* 2009), CNVnator (Abyzov *et al.* 2011) and GenomeSTRiP (Handsaker *et al.* 2011).

The dataset used in the population genomics analysis for the Freeze 2 was a subset of 357,608 non-SNP variants and the complete set of SNP variants (see 2.4.1). Also the high quality second-generation assembly genome of *D. simulans* (Hu *et al.* 2013) was used as outgroup species.

### 2.1.2 Recombination data

**Freeze 1.** The recombination calculator of Fiston-Lavier (Fiston-Lavier *et al.* 2010) was used to estimate the recombination rate in centiMorgans per megabase (cM/Mb) in windows along each chromosome arm or by gene. The recombination rate at the center of each interval was the used value. The calculator is based on Marey maps (Marais *et al.* 2001), where both the genetic (cM) and physical (Mb) positions of 644 genes were fitted to a third-order polynomial curve for each chromosome arm, and the recombination rate at any given physical position estimated as the derivative of the curve (Fiston-Lavier *et al.* 2010).

**Freeze 2.** The recombination values for each interval were estimated from the high resolution recombination map of *D. melanogaster* of Comeron *et al.* (2012) which was obtained in parallel with the development of this thesis. The map was made calculating the crossing over (c) indicated in centimorgans (cM) per megabase (Mb) per female meiosis.

### 2.1.3 Diversity measures

**Nucleotide variation estimates (Freeze 1 phase).** We computed various diversity measures for the whole genome, by chromosome arm (X, 2L, 2R, 3L, 3R), by chromosome region (three regions of equal size in Mb — telomeric, middle and centromeric- were defined) and in 50-kbp non-overlapping windows.

Diversity was estimated as the number of segregating sites ( $S$ ) (Nei 1987), the total minimum number of mutations ( $\eta$ ) (Tajima 1996), the number of singletons, nucleotide diversity ( $\pi$ ) (Nei 1987), Watterson's estimator of nucleotide diversity per site ( $\vartheta$ ) (Nei 1987, Tajima 1993), and the Jukes-Cantor corrected divergence per site ( $k$ ) (Jukes & Cantor 1969). Linkage disequilibrium was estimated as the number of haplotypes ( $h$ ) and haplotype diversity ( $Hd$ ) (Nei 1987), Fu's  $F_s$  statistic (Fu 1997),  $D$  (Lewontin & Kojima 1960), the absolute value of  $D$  ( $|D|$ ),  $D'$  (Lewontin 1964), the absolute value of  $D'$  ( $|D'|$ ), and  $r^2$  (Hill & Robertson 1966, Kelly 1997). The different  $D$ s and the  $r^2$  estimates were computed by averaging over all comparisons of polymorphic sites in a window. Several neutrality tests were applied to the data: Fu & Li's  $D$  and  $F$  statistics (Fu & Li 1993), Fay & Wu's  $H$  statistic (Fay & Wu 2000) and Tajima's  $D$  statistic (Tajima 1989).

**Non-SNP variation estimates (Freeze 2 phase).** We estimated various diversity measures for the whole genome and by chromosome arm (X, 2L, 2R, 3L, 3R, 4) in 100-kb non-overlapping windows.

Aside from re-estimations of  $\pi$  using SNP data from the freeze 2, we have calculated  $\pi$  and divergence for indels ( $\pi_{\text{indel}}$  and  $k_{\text{indel}}$ , see 2.4.3) together with minor allele frequency (MAF) and derived allele frequency (DAF) distributions for indels.

## 2.2 PopDrowser: the Population *Drosophila* Browser

---

### 2.2.1 Selection of GBrowse as a framework for the PopDrowser

Given that our goal in the DGRP project was to carry out the Genome-wide molecular population genetic analyses for the sequenced genomes, a population genome browser was a necessary tool to contain both raw data and the estimated population genetics parameters. At that

moment, no genome browser devoted to population genomic data was available. So, during the Master's Thesis phase of the PhD candidate, we searched and compared current web-based genome browser frameworks to create a population genome browser. At one point we narrowed our options between three candidates: UCSC Genome Browser (Kent *et al.* 2002), Ensembl (Hubbard *et al.* 2002) and GBrowse (Stein *et al.* 2002) (Table 2.1).

These three browsers were, at that moment, the only free and open source genome browser platforms incorporating a web interface and customizable features. Table 2.1 shows the clear impact that the programming language has in the overall performance of the application. UCSC outrival the other two platforms in terms of loading speed. Even though the three platforms can be locally installed, both UCSC and ENSEMBL platforms are tightly developed around the data they currently provide. This was really an issue, although being technically possible to incorporate our own annotation databases into a local UCSC or ENSEMBL installations, their documentation only covered the creation of a mirror installation (exact copies of the browser, with both interface and databases).

At the end, the ability to control all aspects of the browser was the decisive factor. We selected GBrowse (Stein *et al.* 2002) for its *generic* philosophy that suited the most our objective to create a genome browser from scratch. Since Stein's browser was created with portability and flexibility in mind, it had the most complete and accessible installation and configuration options of the three considered systems. GBrowse gives absolute control of every aspect of the browser's administration: from the annotation databases, the basic configuration and functionalities, to the visual aspect of the interface. GBrowse even allows to extend functionalities via a plug-in system for custom scripts.

### 2.2.2 Interface and implementation

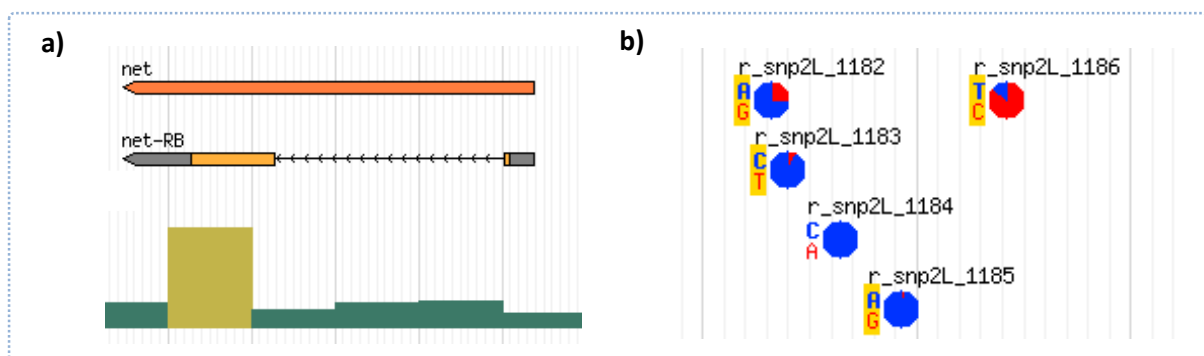
Our browser includes all the default built-in functions of GBrowse to search and display chromosomal regions, select tracks, add custom annotations in standard formats and download data from a particular region. Data is displayed through glyphs, the graphical representations used for annotations in BioPerl (Stajich *et al.* 2002). Each glyph-based annotation is associated to a specific track. An activated track, thus, allows the visualization of the corresponding glyphs (Figure 2.1).

**Table 2.1** Comparison of three candidate genome browser platforms

	ENSEMBL	GBrowse	UCSC	
<b>Programming Language</b>	Perl	Perl	C	
<b>License</b>	Free for academic use (mirror)	Open Source	Free for academic use (mirror)	
<b>Type</b>	Browser + data mining + API	Generic Browser	Browser + data mining + API	
<b>Ease of use (installation / configuration)</b>	Difficult	Very Easy	Very Difficult	
<b>Documentation</b>	Incomplete and confusing	Very Complete	Only mirror installation instructions	
<b>Mailing list / Support</b>	Yes	Yes	No	
<b>Customization possibilities</b>	Only some HTML areas	Config files, Perl source code, configurable glyphs (BioPerl), multiple DDBB, plug-in system, accessible html/css	No	
<b>Maximum zoom</b>	200Kb / 1Mb*	No limit	No limit	
<b>Simple track loading times (seconds):</b>				
	<b>20Kb</b>	12.86	1.71	1.96
	<b>1Mb</b>	13.27**	6.88	2.94
	<b>23Mb</b>	28.05**	7.28	3.78

Track loading times correspond to a single track (genes) with different zoom levels in the chromosome 2L of *D. melanogaster*.

\*Maximum zoom depends on the species genome  
 \*\*ENSEMBL does not display detailed view at this zoom level



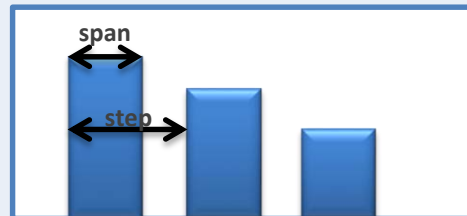
**Figure 2.1** Examples of glyph representations of annotation features in the PopDrowser. (a) GBrowse default glyphs examples: Representation of a gene with the generic rectangle glyph (first in orange) and the 'gene' glyph that represents information of introns and UTRs as well. Below them we have data in histogram, using the xy\_plot glyph. (b) Customized glyph example: SNP frequency information displayed in the PopDrowser with a custom glyph adapted from code of the HapMap Browser.

### Box 3: Common annotation file formats

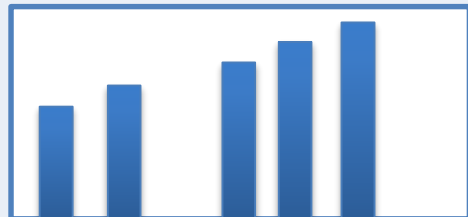
```
##gff-version 3
##sequence-region ctg123 1 1497228
ctg123 . gene 1000 9000 . + . ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000 1012 . + . Parent=gene00001
ctg123 . mRNA 1050 9000 . + . ID=mRNA00001;Parent=gene00001
ctg123 . mRNA 1050 9000 . + . ID=mRNA00002;Parent=gene00001
ctg123 . mRNA 1300 9000 . + . ID=mRNA00003;Parent=gene00001
ctg123 . exon 1300 1500 . + . Parent=mRNA00003
ctg123 . exon 1050 1500 . + . Parent=mRNA00001,mRNA00002
ctg123 . exon 3000 3902 . + . Parent=mRNA00001,mRNA00003
ctg123 . exon 5000 5500 . + . Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon 7000 9000 . + . Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS 1201 1500 . + 0 ID=cds00001;Parent=mRNA00001
ctg123 . CDS 3000 3902 . + 0 ID=cds00001;Parent=mRNA00001
ctg123 . CDS 5000 5500 . + 0 ID=cds00001;Parent=mRNA00001
```

**Figure 2.3 Generic Feature Format file version 3 (GFF 3).** The GFF file is the standard file to use when uploading annotations to the GBrowse, also widely used as standard file for annotations. It is a tabulated file that consists of a set of header lines for metadata, followed by one annotation per line with data distributed in 9 columns: (i) sequence/chromosome ID, (ii) source (free text), (iii) feature type (defined in the Gene Ontology website), (iv) start (1-based), (v) end (1-based), (vi) score, (vii) strand, (viii) phase, (ix) aAttributes. The 9<sup>th</sup> column (attributes) is used to specify relationships between annotations (exons of a gene, for example), and is fully customizable by the user to add any extra information desired. Optionally, at the end of a GFF, the corresponding Fasta sequence can be included. (Example GFF3 from [sequenceontology.org](http://sequenceontology.org))

```
fixedStep chrom=chr19 start=49307401
step=300 span=200
1000
900
800
700
600
...
```



```
variableStep chrom=chr19 span=150
49304701 10.0
49304901 12.5
49305401 15.0
49305601 17.5
49305901 20.0
...
```



**Figure 2.4 WIG format definition.** This type of file is used to store and display huge amounts of quantitative data distributed along the genome in fixed window sizes (defined as 'span'). There are two versions: (i) **Fixed step WIG**, where the distance between windows ('step') is fixed. Since the two values are fixed, there is no need to store coordinates for each annotation, only the quantitative data is stored. (ii) **Variable step WIG**. Here the distance between windows is variable, so at least the start coordinate of the window must be stored along with the quantitative data. Window size remain fixed. (Example WIG from [genome.ucsc.edu](http://genome.ucsc.edu))

### Box 3 (cont): Common annotation file formats

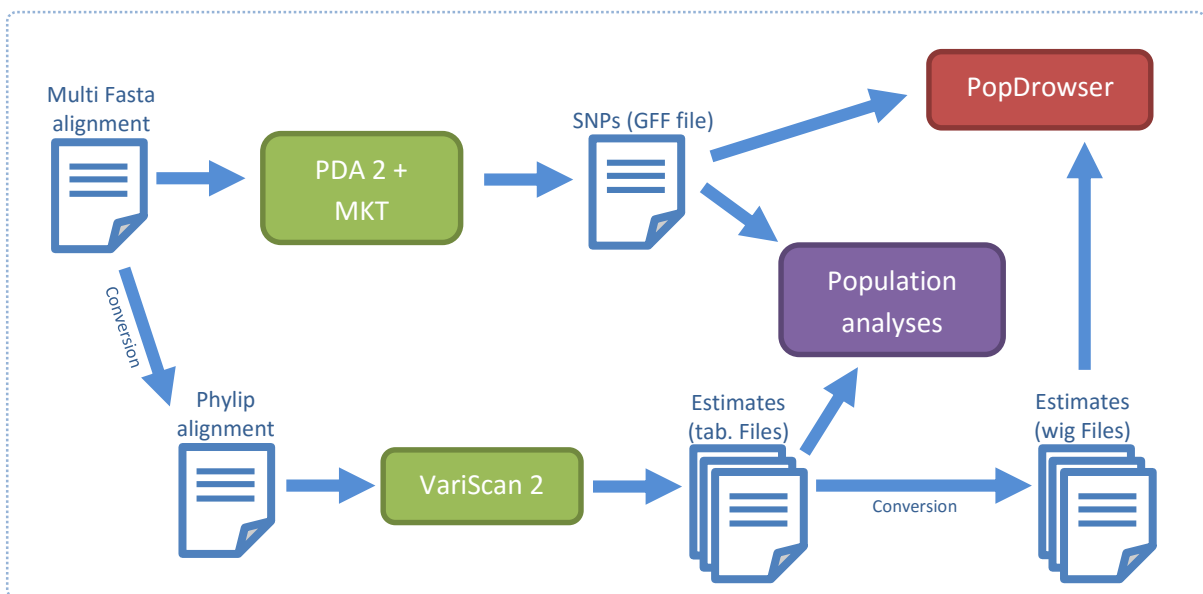
```
##fileformat=VCFv4.1
##filedate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=1, length=62435964, assembly=B36, md5=f126cdf8a6e0c7f379d618ff66beb2da, species="Homo sapiens", taxonomy=x>
##phasing=partial
##INFO=1, Type=Integer, Description="Number of Samples With Data">
##INFO=2, Type=Integer, Description="Total Depth">
##INFO=3, Type=Float, Description="Allele Frequency">
##INFO=4, Type=String, Description="Ancestral Allele">
##INFO=5, Type=Flag, Description="dbSNP membership, build 129">
##FILTER=1, Type=Flag, Description="Quality below 10">
##FILTER=2, Description="Less than 50% of samples have data">
##FORMAT=1, Type=String, Description="Genotype">
##FORMAT=2, Type=Integer, Description="Genotype Quality">
##FORMAT=3, Type=Integer, Description="Read Depth">
##FORMAT=4, Type=String, Description="Haplotype Quality">
##CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0/1:48:1:51,51 1/0:48:8:51,51 1/1:43:5:..
20 17330 . T A 3 Q10 PASS NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0/1:49:3:58,50 0/1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1/2:21:6:23,27 2/1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0/0:54:7:56,60 0/0:48:4:51,51 0/0:61:2
20 1234367 microsatl GTC G,GCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

**Figure 2.5** Variant calling format (VCF). Modification of the GFF format specifically designed to store variation information. It is a tabulated file that consists of a set of header lines for metadata, followed by one annotation per line with data distributed in a variable number of columns with 9 fixed: (i) sequence/chromosome ID, (ii) start position (1-based, see Figure 1.10), (iii) annotation ID, (iv) reference allele, (v) alternative allele, (vi) quality, (vii) filters, (viii) extra information, (ix) format of the individual information. With the 10<sup>th</sup> column, starts specific information about each individual analysed (using the format defined in the 9<sup>th</sup> column). As many columns as individuals can be added after this one. Information in the 8<sup>th</sup> and 9<sup>th</sup> column is highly customizable, and the data types used must be defined in the metadata first. (Example VCF with 3 samples [NA00001, NA00002, NA00003] from 1000genomes.org)

Together with these basic functions, we designed our browser having two great functionalities in mind: (i) First, a large, static and precomputed collection of variation estimates along the genome prepared for fast access; and (ii) the possibility to perform custom re-estimation of population statistics on-the-fly by the user.

In terms of hardware the browser's host server has 2 Intel Xeon 3Ghz processors and 32GB RAM. For the software, the operative system of the current implementation is an Ubuntu 10.04 Linux x64 with the Apache web-server.

**Precomputed estimates.** Along with basic *D. melanogaster* 5.13 annotations, all Freeze-1 population genetics estimates listed in section 2.1.3 are introduced into the browser's databases as precomputed information. This precomputed estimates of several DNA variation measures along each chromosome arm are obtained with the combined implementation of the programs PDA2 (Casillas and Barbadilla 2006), MKT (Egea *et al.* 2008) and VariScan2 (Hutter *et al.* 2006) (Figure 2.2). (see 2.3.1 for details in the computation of these estimates for the population genomics analyses).



**Figure 2.2** Freeze-1 population genomics estimates pipeline.

All summary estimates are computed all along the chromosomes in non-overlapping sliding windows of 50, 100, 500, 1000, 10.000, 50.000 and 100.000 base pairs. All functional annotations are stored with the standard GFF3 format (Box 3, Figure 2.3), then uploaded to



MySQL databases; while most quantitative results are stored in wiggle text format (Box 3, Figure 2.4) and displayed in the browser as boxplots using *wiggle\_xyplot* glyphs.

**On-the-fly estimates.** PopDrowser allows the re-estimation of a selected population genetics measure in any given region of the genome. Thanks to the collaboration with the VariScan 2 developers, a modified version of the software is used to re-calculate any estimate directly from the interface of the PopDrowser. All genome browsers to date are designed to display a single region of the genome, hence, our on-the-fly estimates are available only for a single region at a time as well. For performance issues, a maximum of 1MB per re-estimation was defined. In a similar way, a user can download the aligned sequences from the region in view in the browser to further sequence analyses by using other software outside PopDrowser.

### 2.3 Nucleotide variation description and analysis along the genome of a natural population of *Drosophila melanogaster*

---

We used the DGRP Freeze 1.0 Illumina sequence data and genome sequences from *Drosophila simulans* and *Drosophila yakuba* (Clark *et al.* 2007) to perform genome-wide analyses of polymorphism and divergence, and assess the association of these parameters with genomic features and the recombination landscape.

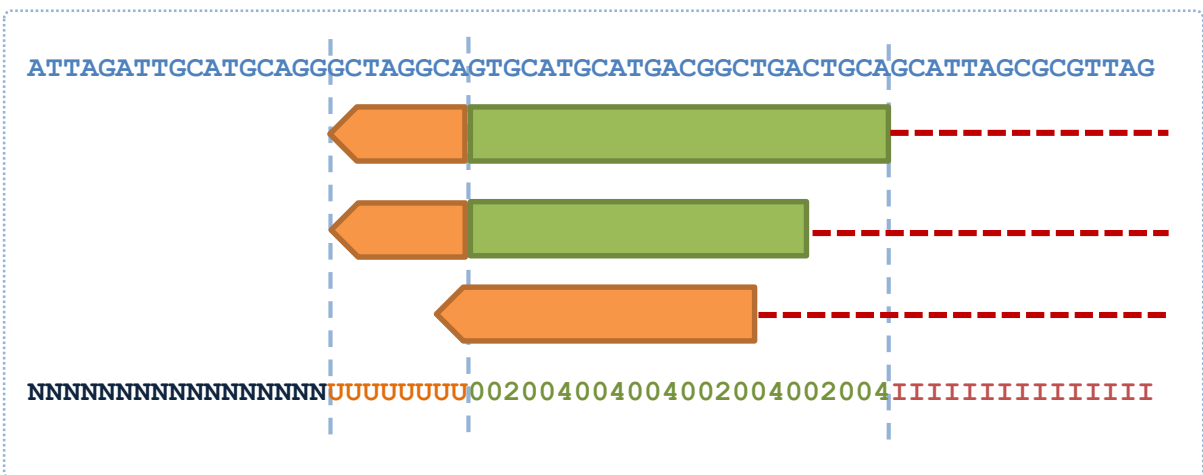
#### 2.3.1 Diversity measures & Linkage disequilibrium

The same Freeze-1 population estimates described in section 2.1.3 and implemented as precomputed tracks in the PopDrowser (see 2.2.2) are used to do the population genomics analysis of the Freeze-1 data. These measures were estimated by implementing PDA2, MKT and VariScan2, the R statistical package (for graphics) and custom Perl scripts (Figure 2.2). Both PDA 2 and VariScan2 can calculate almost the same population genetic estimates, from an initial sequence alignment and either in regions or sliding-windows. However, VariScan2 is coded in C language while PDA2 is a collection of Perl modules. For this reason, VariScan2 can do the estimations much faster than PDA2, which was convenient due to the large number of estimations and window sizes analyzed genome-wide. The DGRP Freeze-1 alignments were converted to Phylip format (Felsenstein 1981) for even better performance of VariScan2. PDA2 and MKT were used to create the scripts in charge of the SNP calling.

Although the quality of the sequences was already remarkable, some filters were applied to the alignments to ensure strength of the population genomics analyses: (i) ambiguous bases were not considered; (ii) the number of lines analyzed in each window was fixed at 140, which minimized the loss of sites in each window while accounting for the bias introduced by the clustering of polymorphic sites with ambiguous nucleotides; in addition, (iii) windows in which 50% or more sites were ambiguous or unaligned were excluded completely.

### 2.3.2 Re-coded whole-genome consensus sequence

For some analyses, it was required to know the functional class for each position in the genome. To do this, a new re-coded sequence was created using a custom Perl script, the reference and gene annotations for *D. melanogaster*. Each position in the genome was annotated in the following categories: non-coding, small introns ( $\leq 100$  bp), long introns ( $> 100$  bp), UTRs and synonymous and non-synonymous coding positions (in the form of 0, 2 or 4 fold degenerate sites). In the cases where multiple annotations overlapped a position we selected a single category following this criterion: 0-fold  $>$  2-fold  $>$  4-fold  $>$  UTR  $>$  Small intron  $>$  long Intron  $>$  Intergenic (Figure 2.6).

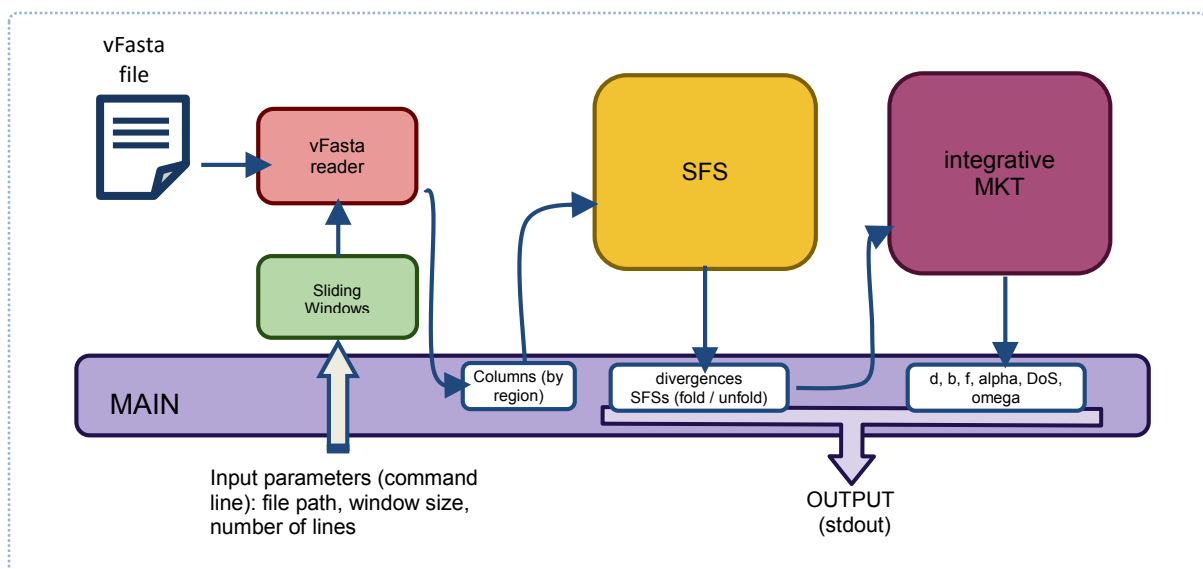


**Figure 2.6** Example of a consensus recoded fragment of the genome with a gene with multiple transcripts. (N) is for intergenic in dark blue, (U) is for UTR in orange, (0, 2, 4) are the degeneracies of the coding regions in green, (I) is for introns in red.

### 2.3.3 Detecting and estimating natural selection

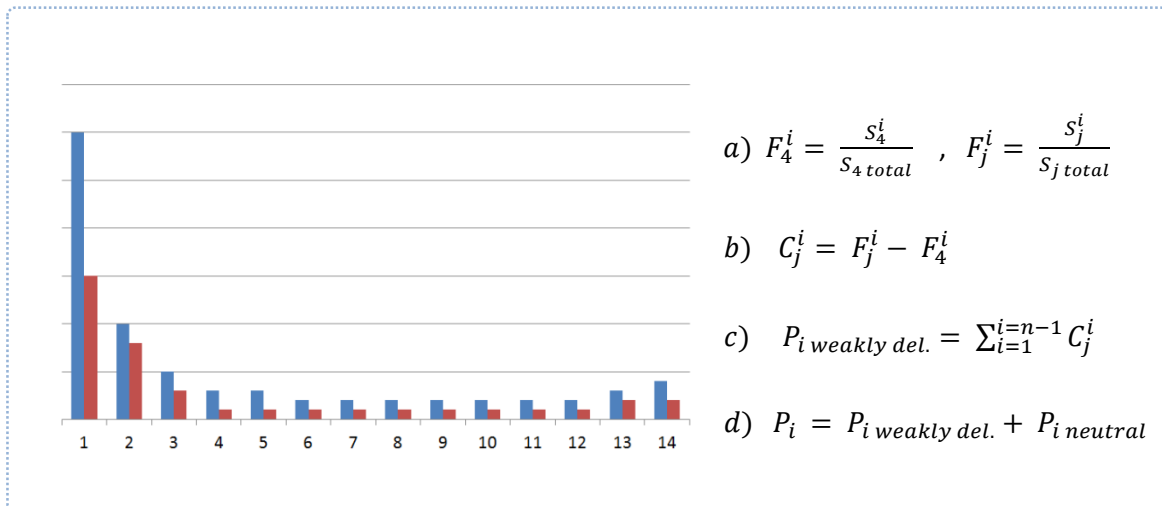
Adaptive mutations and weakly deleterious selection act in opposite directions on the MKT, so if both selection events are occurring, they amount and sign of selection will be underestimated. To take adaptive and slightly deleterious mutation mutually into account, a method that

incorporates site frequency spectrum (SFS) data to the framework of the McDonald & Kreitman (McDonald & Kreitman 1991) test (MKT), the Integrative-MKT (see Box 2) was used to analyze the DGRP freeze 1.0 genome data (Supplementary Material and Methods of Mackay *et al.* 2012). To estimate selection along chromosomes, the integrative-MKT was implemented in software written in Java (Figure 2.7). The re-coded consensus sequence is used in the tests, but separating coding positions between synonymous or non-synonymous instead of the number of degeneracies. In the case of a 2-fold position, it was classified either synonymous or non-synonymous only in the simple cases and not in the complex ones (Kumar *et al.* 2001).



**Figure 2.7** Integrative-MKT Java modules pipeline. Arrows indicate flux of data between the program modules. Input combines a precomputed file (vFasta) and command line parameters. Output is stdout. The program does not create intermediate files.

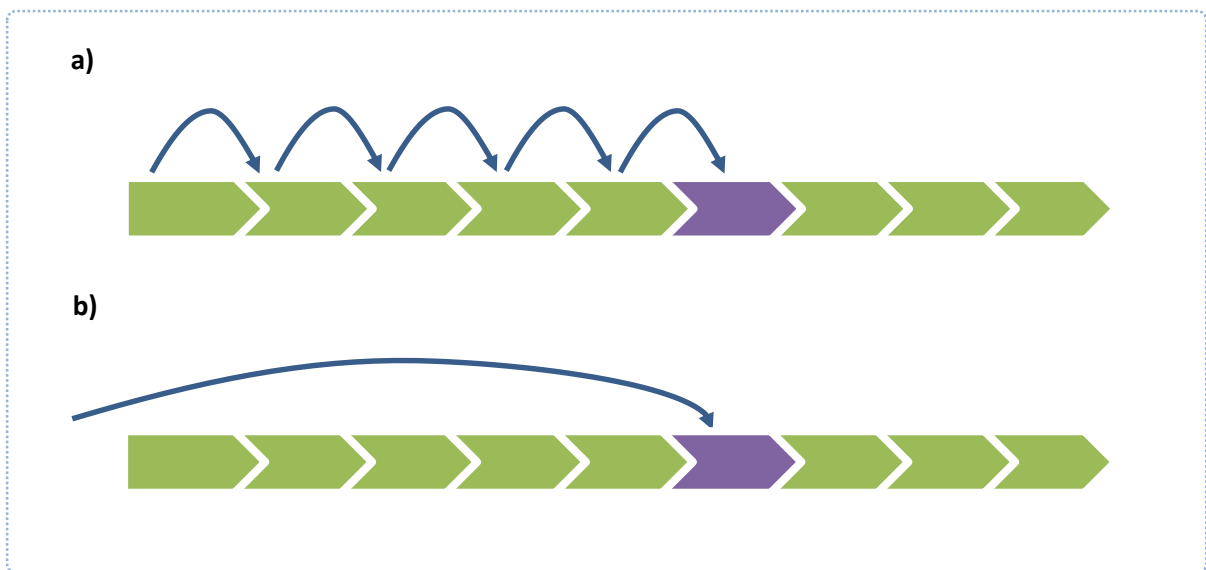
The Integrative-MKT allows estimating five different regimes of selection acting on new mutations for any given region of the genome: the fraction of strongly deleterious ( $d$ ), weakly deleterious ( $b$ ) and neutral sites ( $f$ ) for the given region (sites that have become neutral ( $\gamma$ ) was calculated as a subset of ( $f$ )). It has also been implemented the calculation of the proportion of adaptive fixations ( $\alpha$ ) (Smith & Eyre-Walker 2002), the Direction of Selection ( $DoS$ ) (a weighted and unbiased estimator of the neutrality index, Stoletzki & Eyre-Walker 2011) and the ratio  $\omega_\alpha$  (the rate of adaptive evolution relative to neutral evolution, Gossman *et al.* 2012). To account for the fraction of neutrally segregating sites in the selected class (see Box 2) we use the information of the SFS as explained in figure 2.8.



**Figure 2.8** Calculation of the fraction of neutrally segregating sites in the putatively selected class by comparing the SFSs. As seen in Box 2,  $P_i$  (the count of segregating sites in the selected class of the standard MKT table) must be decomposed between neutral variants ( $P_{i\ neutral}$ ) and weakly deleterious variants ( $P_{i\ weakly\ del.}$ ). The site frequency spectrum of the selected class (blue) is compared with the spectrum of the 4-fold degenerated coding sites (the selected neutral class in the integrative-MKT software). First, (a) each value in the SFS is divided by the total number of segregating sites of its class. Then, (b) each value of the 4 fold divided SFS is subtracted from the selected SFS class. All resulting values are (c) added together, this is the fraction of weakly deleterious sites in the selected class, finally, (d) the fraction of neutral sites is easily inferred.

**Optimizing file access.** An unstoppable trend in genomics is the exponential grow in the amount of data available. This usually translates into bigger files. To tackle the impact of huge volume of data on the performance of genomic and bioinformatics analyses many strategies can be considered, for instance the usage of more optimized programming (using lower level languages like C, not at hand for every bioinformatician) or optimizing the reading (access) of the files. In the case of the access to a file, in a classical way, this is sequential. This means that if we have interest in data at the end of the file, the sequential program will move through every record before accessing our data of interest; increasing the processing time (Figure 2.9a). It is possible to access precise data without reading the whole file, this is called **direct access** (or random access, Figure 2.9b). There are two main direct access strategies: (i) the creation of indexes, companion files that store the position/bit where some landmarks are stored in our file to avoid reading the whole file; and (ii) files with fixed number of characters per row (this means that each line will have the same number of bits), this allows us to infer the bit our row is inside the file and access it directly.

We wanted to provide our software with direct access capabilities, and since the creation of index files was out of the expertise of the PhD. candidate, we decided for the fixed character number strategy. Our data of interest was the nucleotide at each position for each line, so the normal Multi-Fasta alignment files were converted into a new vertical Multi-Fasta format (named vFasta), where each position was a row and each individual genome was a column. As we had a fixed number of genomes, this implied a fixed number of characters per row in the vFasta file, so direct access techniques could be applied to speed-up the reading of the alignment.



**Figure 2.9** Diagram picturing (a) sequential access and (b) direct access to information inside a file.

## 2.4 Indel variation landscape in the genome of a natural population of *D. melanogaster*

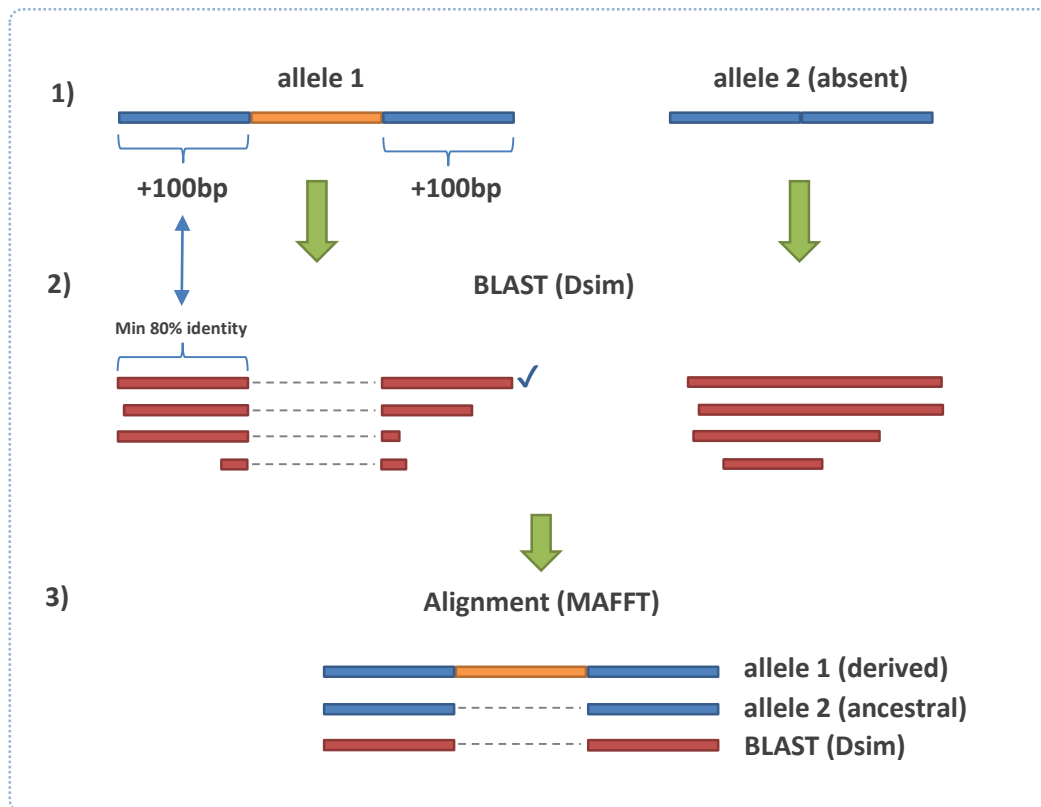
---

### 2.4.1 Filtering structural variants

We used 357,708 JGIL-filtered, biallelic indels present in at least 101 lines to conduct the indel population genomics analyses. We assigned indels to one of six functional classes (coding sequence, 5' and 3' UTR, long [ $>100$  bp] and short [ $\leq 100$  bp] introns, intergenic sequence) using the 5.49 version annotations of the *D. melanogaster* reference genome (Marygold *et al.* 2013). We discarded indels spanning more than one functional class, leaving 357,608 indels with a valid functional class. Variant calls for the Freeze-2 were provided in VCF format (Box 3, Figure 2.5).

## 2.4.2 Inferring the ancestral state of Indels

We analyzed insertions and deletions separately, after first polarizing ancestral and derived states with respect to the high quality second-generation assembly genome of *D. simulans* (Hu *et al.* 2013) as an outgroup. We inferred the derived allele status for 210,268 indels.



**Figure 2.10** Protocol to infer the ancestral state of an indel (indel polarization). (1) We add 100 nucleotides flanking each side of each indel allele (blue). Since we only have bi-allelic indels, all indels are a combination of presence (orange)/absence (nothing) of sequence. (2) Both extended alleles are BLASTed to the genome of an outgroup species (*D. simulans*). We select the longest BLAST result, taking into account a minimum 80% identity for the added flanking sequence. (3) We align the 2 extended alleles with the selected blast result. The allele with most identity with the outgroup sequence is selected as ancestral and the other allele as derived.

We followed a strict protocol to polarize indel alleles (Figure 2.10). (i) For each biallelic indel, we added 100 nucleotides 5' and 3' from the reference sequence. (ii) We did a BLAST (Altschul *et al.* 1990, Camacho *et al.* 2009) search for both allelic sequences to *D. simulans*, retaining the longest *D. simulans* sequence for the next step. We discarded blast results with multiple hits and required that a valid hit must include at least 80% of the added nucleotides in step (i). Because larger indels do not always result in a valid blast alignment, we required valid blast hits for both

alleles for indels  $\leq 25$  bp, while for indels  $> 25$  bp only one valid hit for any of the two alleles was considered sufficient. (iii) We simultaneously aligned both indel allele sequences plus the corresponding *D. simulans* sequence using MAFFT software (Katoh and Standley, 2013) using the '--globalpair --maxiterate 20' options. This gave comparable results compared with other alignment software such as ClustalOmega (Sievers *et al.*, 2011), TCOFFEE (Notredame *et al.*, 2000) and MUSCLE (Edgar, 2004). (iv) We trimmed the alignment at 25 bp before and after the indel initial and end coordinates, respectively. We assigned derived allele status to the allele sequence which differs from the *D. simulans* sequence in this alignment region. From the trimmed alignment we discarded any fixed indels between *D. simulans* and *D. melanogaster*, or any partially overlapping gap. (v) We determined insertion or deletion status based on whether the derived allele sequence adds or removes nucleotides compared with *D. simulans*.

We manually checked a random sample of 500 derived indels to which our polarizing protocol was applied; all were correct. Therefore, we conclude that the specificity of our procedure is very high, although we excluded 41% of the original indel data set from our evolutionary analyses.

### 2.4.3 Calculating Indel Variation

We have used a measure analogous to nucleotide diversity ( $p$ ) to describe indel polymorphism,  $\pi_{indel}$ . Basically we have considered every indel as biallelic and did not take into account its size. For biallelic SNP or indel events we can calculate their intrinsic variation or  $k$ , basically, an estimate like  $\pi$  (Nei 1987) for a single indel/SNP and assuming only 2 alleles:

$$k = \frac{\text{Freq. allele A} * \text{Freq. allele B}}{\binom{n}{2}}$$

Then nucleotide diversity ( $\pi$ ) for a biallelic SNP or ( $\pi_{indel}$ ) for biallelic indels can be simplified as the sum of each intrinsic variation in a given genomic region  $m$  (in number of nucleotides):

$$\pi = \frac{\sum_{i=1}^n k_i}{m}$$

We used an analogous measure to the SNP divergence to estimate divergence in indels ( $k_{indel}$ ) (Librado and Rozas 2009). We estimated fixed and polarized indel divergence for the *D. melanogaster*, *D. simulans*, and *D. yakuba* lineages using the multiple alignment *caf1\_6way* (*D. melanogaster*, *D. simulans*, *D. yakuba*, *D. pseudoobscura*, *D. ananassae*, *D. erecta*) from the

VISTA Browser (Frazer *et al.* 2004). We used this alignment because was the only one we found that has the three species of interest (*D. melanogaster*, *D. simulans*, and *D. yakuba*) aligned at the same time, the other three species are not used.

We estimated these diversity measures for the whole genome and by chromosome arm (X, 2L, 2R, 3L, 3R, 4) in 100-kb no-overlapping windows. We also estimated the minor allele frequency (MAF) distribution for indels and the derived allele frequency (DAF) distributions for both deletions and insertions. It's generally assumed that synonymous SNPs segregate neutrally, however some studies suggest that SNPs in small introns (< 120bp) could act even more neutrally (Parsch *et al.* 2010). Thus, frequencies of synonymous SNPs and SNPs in small introns are considered in the MAF and DAF distributions as putative neutral classes. We chose our threshold in introns 100bp long or less as small introns, since we empirically observed a clear clustering in the genome between introns below and above this size.

We used the nonparametric Spearman rank correlation coefficient (rho or  $\rho$ ) to test for covariation among the diversity estimates. We used the high resolution recombination map of *D. melanogaster* by Comeron *et al.* (2012) to correlate recombination with the estimated diversity measures.

#### 2.4.4 Estimating the proportion of adaptive fixations in indels

With DFE-alpha (Eyre-Walker & Keightley, 2009, see 1.1.7) we tried to estimate the effects of adaptive selection on indels. Similar to the MK test, DFE-alpha uses information from two class sites, one putatively neutral and another to be tested for adaptive selection, but instead of using counts of variants it uses the site frequency spectrum of each class along with divergence data. DFE-alpha estimates the distribution of fitness effects and reports the proportion of adaptive substitutions ( $\alpha$ ), and the relative rate of adaptive substitutions relative to the neutral substitutions ( $\omega_\alpha$ ). If we observe the formula to estimate  $\alpha$  in the MK test:

$$\alpha = 1 - \frac{D_s P_i}{D_i P_s}$$

And compare with the method to calculate alpha in the DFE-alpha:

$$\alpha = \frac{d_N - d_S \int_0^\infty 2Nu(N, s) f(s|a, b) ds}{d_N}$$



We can see that the main addition is the consideration of the effects of slightly deleterious mutations ( $f(s|a, b)$ ; a gamma distribution of scale parameter  $a$  and shape parameter  $b$ ).  $2Nu(N, s)$  refers to the probability of fixation of a new mutation with selective effect  $s$  and population size  $N$ .  $d_N$  and  $d_S$  are the number of selected and neutral substitutions per site. The complete numerator represents the difference between observed and expected rates of selected substitutions no assuming adaptive fixation.

Since it is not clear in which functional class indels are the most neutral, we used intergenic, small and long intron indels as putatively neutral classes with DFE-alpha. We assume these three classes of sites to be the less functionally and in this way we expect to clarify which of them is the best candidate for neutral indel class with the results (see Discussion).

## 2.5 Summary of used and developed software

---

The PopDrowser is based in GBrowse (Stein *et al.* 2002). The precomputed estimates of several DNA variation measures are obtained using VariScan 2 (Hutter *et al.* 2006), PDA2 (Casillas and Barbadilla 2006), MKT (Egea *et al.* 2008), the recombination calculator of Fiston-Lavier (Fiston-Lavier *et al.* 2010); plus *ad-hoc* Perl and Bash scripts to run the other software and parse or convert results. Scripts in the R statistical language were used to plot graphics and run statistics.

The integrative-MKT was developed in Java. The script to create the recoded sequence and the vertical Fasta conversion were developed in Perl.

The indel polarization was done using BLAST (Altschul *et al.* 1990, Camacho *et al.* 2009) and MAFFT (Kato and Standley, 2013). The estimation of natural selection,  $\alpha$ ,  $\omega_a$  y  $\omega_d$  was done using *DEF-alpha* (Eyre-Walker and Keightley 2009). Automation, intermediate steps, file conversion and results parsing was done with *ad-hoc* Python scripts. Most of the indels statistical analysis plus graphic plotting was done with R scripts.

Software and most *ad-hoc* scripts developed for this work are publicly available at <https://github.com/mikyatope/thesis> .



---

## **Part 3.**

### **Results**

---

## 3. Results

### 3.1 PopDrowser: the *Drosophila* Genome Variation Browser

---

The implementation of the DGRP polymorphism map using the GBrowse framework was called PopDrowser and it is freely accessible from <http://PopDrowser.uab.es>. The following is a description of all its implemented tracks and functions.

#### 3.1.1 Output

*Reference annotations.* We have incorporated *D. melanogaster* reference annotations (build 5.13) (Smith *et al.* 2007) from Flybase. These include gene information (mRNAs, CDSs, 6-frame translations), noncoding RNAs, tRNAs, and insertion sites of transposable elements. The last version of the phastCons conservation track (Siepel *et al.* 2005) from the UCSC is also displayed by using the Distributed Annotation System (DAS) protocol (Dowell *et al.* 2001), as well as a track showing the GC content of the reference sequence or nucleotide sequence when the region is zoomed in.

*Recombination estimates.* We have used the recombination calculator of Fiston-Lavier (Fiston-Lavier *et al.* 2010) to estimate the recombination rate by megabase (*cM/Mb*) in windows along each chromosome arm or in specific gene coordinates. The calculator is based on Marais maps (Marais *et al.* 2001, see 2.1.2). For the representation, we only considered the rate at the central point of the interval.

*Density tracks.* The density of some genomic features has been calculated in sliding windows of 10, 50 and 100Kb along the genome. We have density tracks from reference features (genes, microsatellites, transposable elements and coding sequence) and DGRP features (SNPs and sequencing errors).

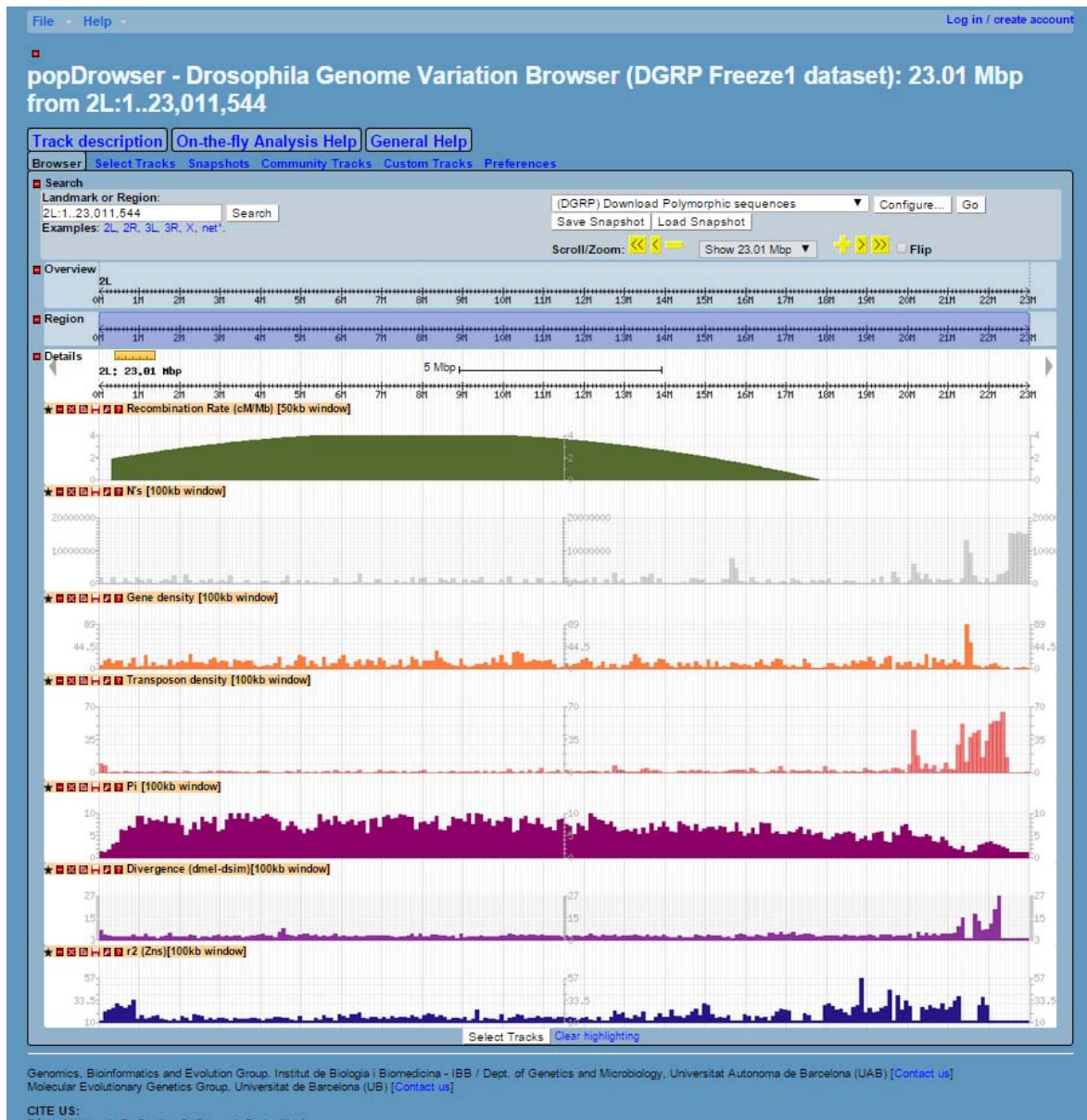
*Nucleotide variants.* The location of nucleotide variants which are polymorphic in the DGRP population (SNPs) or fixed between the *D. melanogaster* reference genome and its phylogenetically closest species *D. simulans* (SNFs), together with their frequencies, are obtained with *ad-hoc* scripts based on the source code from the PDA software (Casillas and Barbadilla 2006) and displayed in the browser. Singletons (segregating sites where the minor allele occurs only once in the sample) are shown with a lighter shade. When a region is zoomed in, the allele frequency of each SNP is displayed as a pie graph glyph that has been adapted from the *allele\_pie\_multi* glyph from HapMap (Frazer *et al.* 2007) and that displays two allele frequencies: the frequency of the major allele and the added frequency of all other alleles (precise frequencies for all alleles can be seen by hovering the mouse over the glyph).

*Summary measures of nucleotide variation.* Precomputed estimations for the main population genetics statistics were obtained using VariScan2 (Hutter *et al.* 2006; see 2.3.1). These statistics include the number of segregating sites ( $S$ ) (Nei 1987), the total minimum number of mutations ( $\tau$ ) (Tajima 1996), the number of singletons, the nucleotide diversity ( $\pi$ ) (Nei 1987), the Watterson's estimator of nucleotide diversity per site ( $\theta$ ) (Nei 1987, Tajima 1993), and the Jukes-Cantor corrected divergence per site ( $K$ ) (Jukes and Cantor 1969).

*Measures of linkage disequilibrium.* Similarly, several measures of linkage disequilibrium are computed using VariScan2. These measures include the number of haplotypes ( $h$ ) (Nei 1987), the haplotype diversity ( $Hd$ ) (Nei 1987), and the Fu's  $F_S$  statistic (Fu 1997). The  $D$  value (Lewontin and Kojima 1960), the absolute  $D$  ( $|D|$ ),  $D'$  (Lewontin 1964), the absolute  $D'$  ( $|D'|$ ), and  $r^2$  (Hill and Robertson, 1968, Kelly 1997) have been computed here by averaging over all comparisons of polymorphic sites in the window.

*Neutrality tests.* Several neutrality tests are also performed using VariScan2. These tests include the Fu & Li's  $D$  and  $F$  (Fu and Li 1993), the Fay & Wu's  $H$  statistic (Fay and Wu 2000), and the Tajima's  $D$  statistic (Tajima 1989). Results of the generalized and the integrative McDonald & Kreitman (MK) tests (McDonald and Kreitman 1991, Egea, Casillas & Barbadilla 2007, Mackay *et al.* 2012; see Box 2) per each gene estimated from other members of our lab are also displayed in the browser. We have a track displaying: the generalized MKT results, the proportion of base substitutions fixed by natural selection ( $\alpha$ ) (Charlesworth 1994), the neutrality index (NI) (Rand *et al.* 1996), the direction of selection (DoS) (Stoletzki and Eyre-Walker 2011), the integrative MKT results (Mackay *et al.* 2012) and MAF and DAF spectra. All estimates have their statistical

significances. The genes in this track are colored depending in their DoS value and significance (grey: non-significant DoS, light green: significant positive DoS, dark green: significant negative DoS).



**Figure 3.1** Initial default page of the Popdrowser displaying a whole view of the 2L chromosome arm with several tracks: Recombination rate in 50kb windows (green), Sequencing errors in 100kb windows (grey), Gene density in 100kb windows (orange), Transposon density in 100kb windows, Nucleotide diversity in 100kb windows (purple), Divergence in 100kb windows (violet), Linkage disequilibrium in 100kb windows (blue).

### 3.1.2 Custom analyses on-the-fly

A powerful and innovative capability of this browser is that it allows performing custom analyses on-the-fly in any genomic region up to 1Mb in size. Once a chromosome region and a certain track have been selected for display, the user can choose to reanalyze that region directly from the browser with custom input parameters (e.g. window and step size, include/exclude singletons, include/exclude missing/gapped sites, etc.). Furthermore, users can choose to either visualize the output of their analyses graphically in the browser as a new track or to download it in a tabulated text file (Figure 3.3).

Another functionality implemented in the browser is the option for users to download the aligned 158 DGRP freeze 1.0 genomic sequences of the region visualized at that moment. Both on-the-fly and sequence downloads have been implemented using the GBrowse plug-in system that, using Perl scripting, allowed us to create a layer to communicate GBrowse with Variscan2 (for the on-the-fly analysis) and the aligned DGRP sequence files.

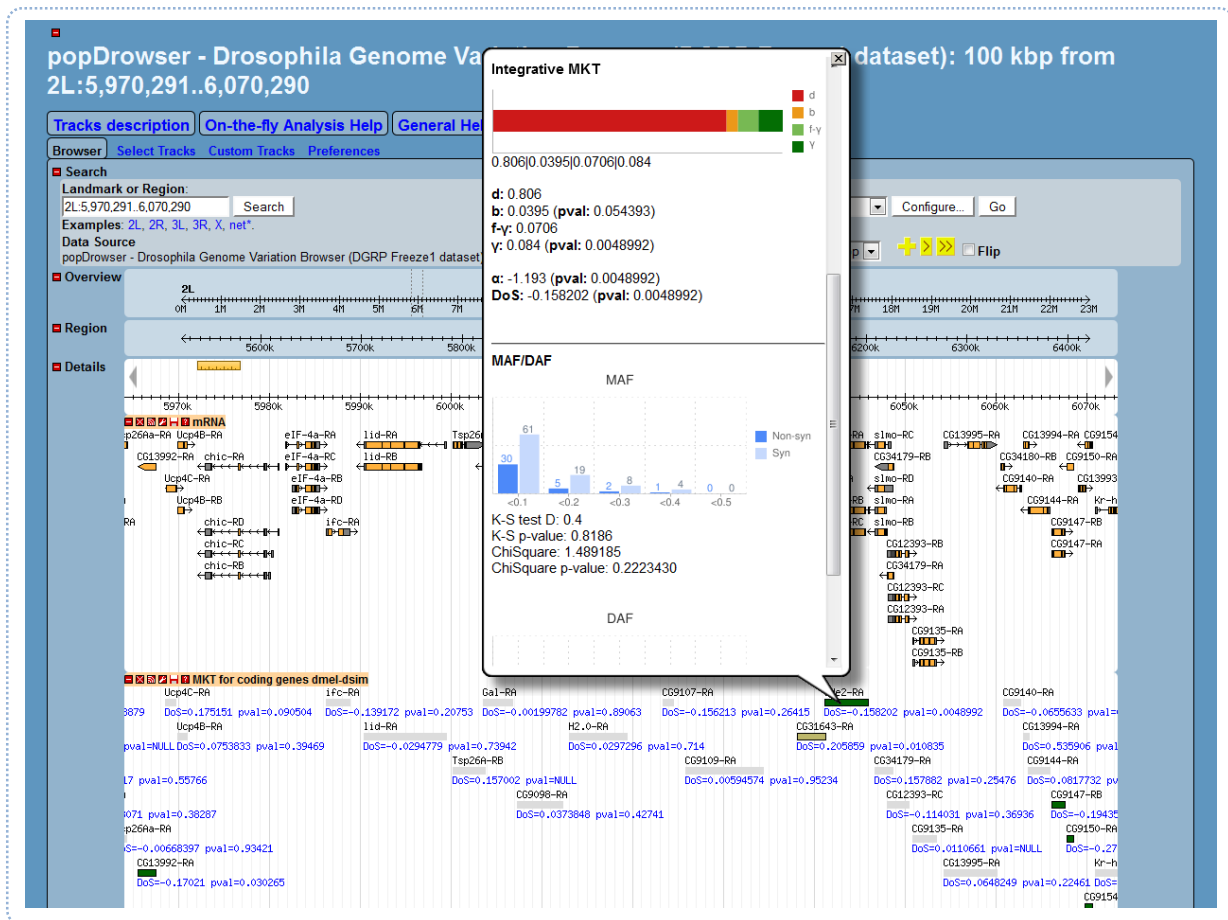
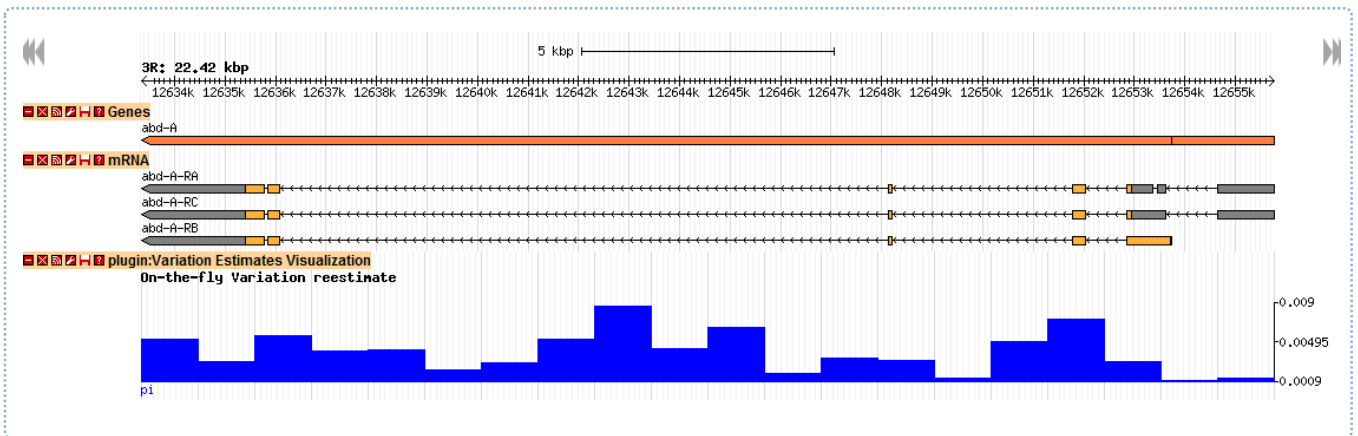


Fig. 3.2 PopDrowser snapshot showing the reference genes track, the standard MKT track and in a popup the results of the integrative MKT, DoS, MAF & DAF specifically for the *ade2-RA* gene.



**Figure 3.3** On-the-fly re-estimation of the nucleotide diversity in 1Kb windows for a selected genomic region in the PopDrowser (blue track). The same information in the blue track can also be downloaded in tabulated text format.

## 3.2 Nucleic variation analysis of a natural population of *Drosophila melanogaster*

### 3.2.1 Polymorphism and Divergence in the chromosome arms of *D. melanogaster*

The nucleotide polymorphism averaged over the entire genome,  $\pi = 0.0056$  and  $\vartheta = 0.0067$ , was similar to previous estimates based on smaller data sets from North American populations (Sackton *et al.* 2009, Andolfatto & Przeworski *et al.* 2001). Average polymorphism on the X chromosome ( $\pi_X = 0.0040$ ) is reduced relative to the autosomes ( $\pi_A = 0.0060$ ) ( $X/A$  ratio = 0.67, Wilcoxon test  $P = < 10^{-16}$ ), even after correcting for the  $X/A$  effective population size ( $4/3 X = 0.0054$ , Wilcoxon test  $P < 0.00002$ ; Table 3.1) since there is only one copy of an X chromosome for 2 copies of each autosome in every cell.

Nucleotide diversity in non-overlapping 50kb windows is shown in figure 3.4a. We can observe a clear pattern of reduction of diversity around the centromeric regions of the autosomes compared with a relatively constant diversity along the X chromosome. Autosomal nucleotide diversity is reduced on average 2.4-fold in centromeric regions relative to non-centromeric regions, and at the telomeres as well. While the reduction of polymorphism in the centromeric region is gradual, affecting the third fraction of the chromosome arm spanning the centromere, the reduction in the telomere is abrupt. Arms of chromosomes 2 and 3 share this pattern, however the reduction of diversity seems to be even more pronounced in the centromere of

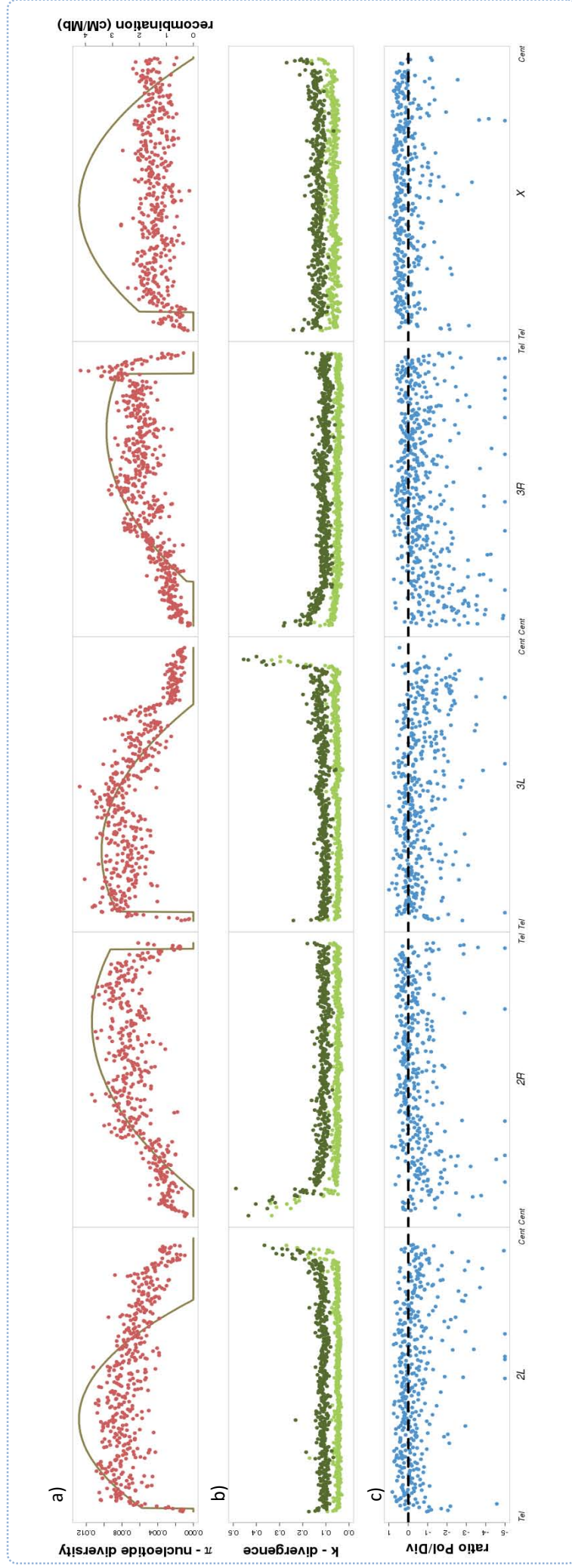


chromosome 3. Chromosome X behaves differentially to autosomes, showing a slight reduction of diversity in the telomeric region that rapidly stabilizes to more uniform levels.

Divergence is more uniform (coefficient of variation,  $CV_k = 0.2841$ ) across chromosome arms than polymorphism ( $CV_\pi = 0.4265$ ; Fig. 3.4b). The observed peaks of divergence near the centromeres could be attributable to several causes: a reduced quality of alignments in these regions producing more sequence errors, higher mutation rates in those regions or higher fixation of slightly deleterious mutations due to low recombination reducing the efficiency of selection (see Discussion). Overall patterns of divergence are similar regardless of the outgroup species used.

### 3.2.2 Recombination landscape

Evolutionary models of hitchhiking and background selection (Begun & Aquadro 1992, Charlesworth *et al.* 1993) predict a positive correlation between polymorphism and recombination rate. Observing figure 3.4a we can see that there is a pattern of less recombination near the centromeric and telomeric regions, quite in parallel to the pattern observed with nucleotide diversity, thus expecting a correlation between the two estimates. This expectation is true only in regions where recombination is less than  $2 \text{ cM Mb}^{-1}$  (Spearman's  $\rho = 0.471$ ,  $P = 0$ ), but recombination and polymorphism behave independently in regions where recombination exceeds  $2 \text{ cM Mb}^{-1}$  (Spearman's  $\rho = -0.0044$ ,  $P = 0.987$ ). The average rate of recombination of the X chromosome ( $2.9 \text{ cM Mb}^{-1}$ ) is greater than that of autosomes ( $2.1 \text{ cM Mb}^{-1}$ ), which may account for the low overall X-linked correlation between recombination rate and  $\pi$ . The lack of positive correlation between recombination and divergence (Table 3.2) excludes mutation associated with recombination as a main cause of the correlation (see Discussion).



**Figure 3.4** Pattern of polymorphism, divergence,  $\alpha$  and recombination rate along chromosome arms in non-overlapping 50-kbp windows.

**a**, Nucleotide polymorphism ( $\pi$ ). The solid curves give the recombination rate (cM Mb<sup>-1</sup>). **b**, Divergence ( $k$ ) for *D. simulans* (light green) and *D. yakuba* (dark green). **c**, Polymorphism to divergence ratio (Pol/Div), estimated as  $1 - [(\pi_{0\text{-fold}}/\pi_{4\text{-fold}})/(k_{0\text{-fold}}/k_{4\text{-fold}})]$ . An excess of 0-fold divergence relative to polymorphism ( $k_{0\text{-fold}}/k_{4\text{-fold}} > (\pi_{0\text{-fold}}/\pi_{4\text{-fold}})$ ) is interpreted as adaptive fixation whereas an excess of 0-fold polymorphism relative to divergence ( $\pi_{0\text{-fold}}/\pi_{4\text{-fold}} > (k_{0\text{-fold}}/k_{4\text{-fold}})$ ) indicates that weakly deleterious or nearly neutral mutations are segregating in the population.

**Table 3.1** Estimates of nucleotide polymorphism ( $\pi$ ), Watterson's  $\vartheta$ , and divergence ( $k$ ) for the whole genome, for each chromosome arm and for regions within arms (based on 50kbp non-overlapping windows). Outgroup species for divergence estimates are *D. simulans* (*dsim*) and *D. yakuba* (*dyak*).

	N	$\pi$			$\vartheta$		
		mean	median	sd	mean	median	sd
<b>ALL</b>	2383	0.0056	0.0059	0.0024	0.0067	0.0073	0.0023
<b>2L</b>	461	0.0068	0.0070	0.0022	0.0080	0.0092	0.0021
<b>2R</b>	423	0.0061	0.0064	0.0024	0.0071	0.0074	0.0022
<b>3L</b>	491	0.0061	0.0068	0.0028	0.0074	0.0081	0.0025
<b>3R</b>	559	0.0051	0.0053	0.0022	0.0063	0.0065	0.0021
<b>X</b>	449	0.0040	0.0043	0.0014	0.0049	0.0053	0.0012
<b>Autosomes</b>	1934	0.0060	0.0064	0.0025	0.0072	0.0078	0.0023
<b>(4/3)X</b>	449	0.0054	0.0057	0.0018	0.0066	0.0070	0.0016

	N	<i>k dsim</i>			<i>k dyak</i>		
		mean	median	sd	mean	median	sd
<b>ALL</b>	2383	0.0620	0.0555	0.0317	0.1283	0.1198	0.0447
<b>2L</b>	461	0.0592	0.0542	0.0238	0.1279	0.1209	0.0389
<b>2R</b>	423	0.0660	0.0538	0.0535	0.1318	0.1163	0.0649
<b>3L</b>	491	0.0597	0.0538	0.0327	0.1267	0.1173	0.0504
<b>3R</b>	559	0.0546	0.0527	0.0142	0.1183	0.1131	0.0334
<b>X</b>	449	0.0729	0.0693	0.0216	0.1403	0.1382	0.0264
<b>Autosomes</b>	1934	0.0594	0.0536	0.0331	0.1256	0.1164	0.0475

**Table 3.2** Spearman and Pearson correlations between nucleotide diversity ( $\pi$ ) / divergence ( $k$ ) and recombination for the whole genome, for each chromosome arm and for regions within arms on 50 kbp non-overlapping windows.

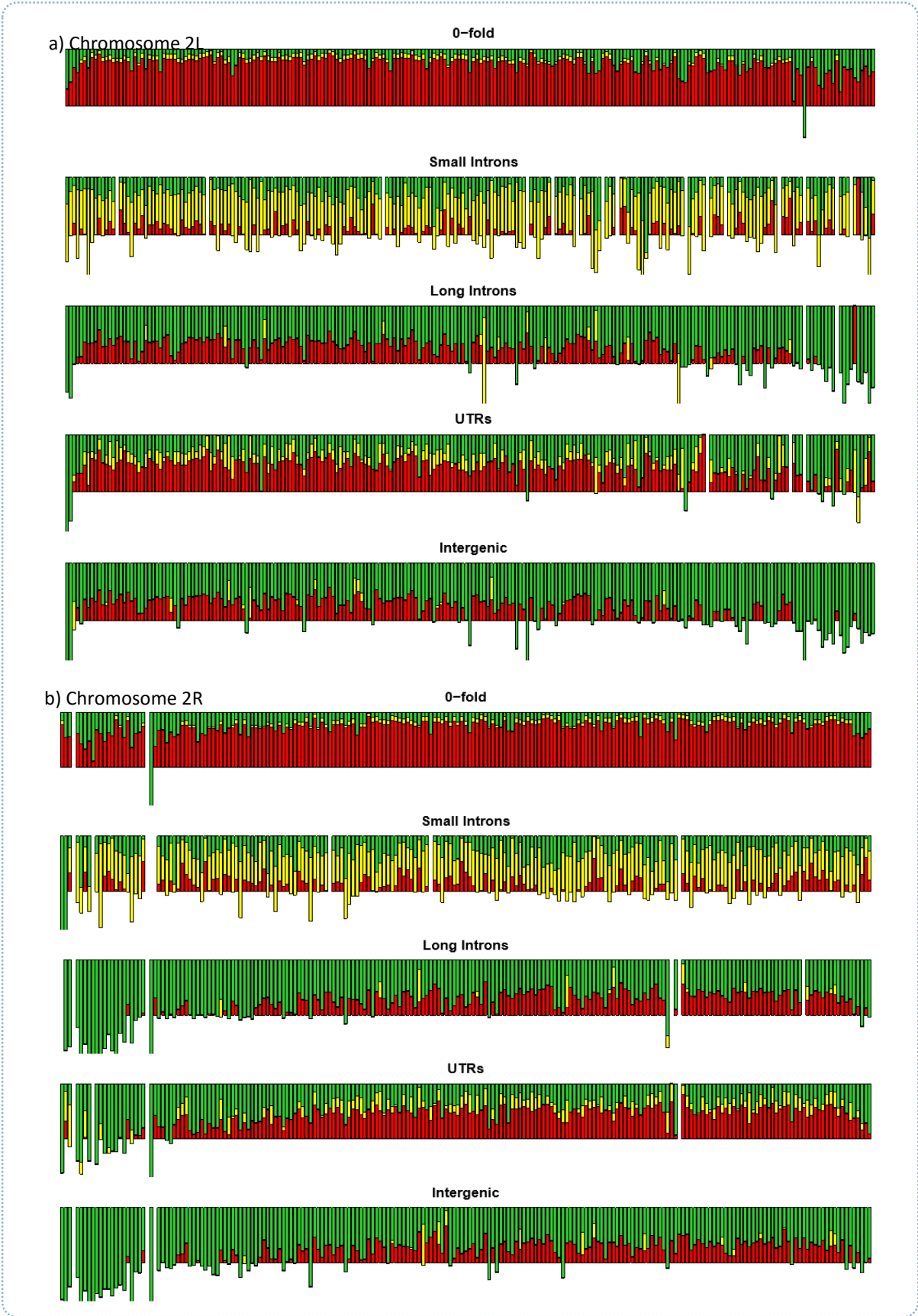
	Correlation between $\pi$ and recombination				Correlation between $k_{dyak}$ and recombination			
	Spearman $\rho$	p-value	Pearson $r$	p-value	Spearman $\rho$	p-value	Pearson $r$	p-value
<b>2L</b>	0.6599	0.0000	0.6952	0.0000	-0.2755	0.0000	-0.3816	0.0000
<b>2R</b>	0.6529	0.0000	0.7598	0.0000	-0.4402	0.0000	-0.5243	0.0000
<b>3L</b>	0.6629	0.0000	0.7745	0.0000	-0.4844	0.0000	-0.4100	0.0000
<b>3R</b>	0.5123	0.0000	0.5660	0.0000	-0.3328	0.0000	-0.3895	0.0000
<b>X</b>	0.1080	0.0239	0.2081	0.0000	-0.1784	0.0002	-0.3204	0.0000

### 3.2.3 Mapping selection across the genome

In figure 3.4c selection is mapped in a genome-wide scale using an estimate analogous to the MKT's neutrality index (see section 1.1.7). Instead of the number of segregating sites, we use the average  $\pi_0$ -fold and  $\pi_4$ -fold values in non-overlapping windows of 50kb. At a first glance, it can be seen that events of either adaptive or deleterious selections seem to occur all along the genome. It also seems to be a reduction of amount of positive selection around the centromeres of the autosomes, which is not visible in the X chromosome. In fact, the X chromosome seems to undergo overall more positive selection than negative selection events.

Trying to improve this initial genome-wide estimate of selection, we implemented the Integrative MKT method into an algorithm to calculate the fraction of neutral, strongly and weakly deleterious sites (see Box 2 and section 2.3.3). In the figure 3.5, we have the results of the genome-wide integrative MKT first implementation for the autosome arms and chromosome X. We can observe that the footprint of natural selection is pervasive all along the *D. melanogaster* genome, although the proportion of sites under the different selection regimes depends on the genomic region and the functional class. In autosomes, there is a higher fraction of neutral alleles (or a lower fraction of selected alleles, particularly strongly selected alleles) in regions near the centromeres (Figure 3.5a) for all functional classes except short introns. This is expected given that short introns are thought to evolve neutrally. In contrast, in the X chromosome there are no evidences for a lower efficacy of natural selection for genes close to the centromere. Moreover, globally the proportion of sites under selection (strongly and weakly) is higher for the X chromosome than autosomes, this is also observed in the analysis gene by gene of the freeze 1 data made also in our lab (Mackay *et al.* 2012). In the gene by gene approach a test was performed comparing the direction of selection (DoS) of genes in low and high recombination areas between autosomes and X chromosome. The DoS comparison indicates lower efficiency of selection for genes in centromeric regions in autosomes (Mackay *et al.* 2012). Altogether, our results suggest a greater efficiency of natural selection in the X chromosome relative to autosomes (Figure 3.5b).

Different gene functional regions, or site classes, show different proportion of sites under different selection regimes. Non-synonymous sites show a greater evidence of strongly deleterious selection than the other classes and very little, but uniform along the genome, proportion of sites under weak (negative) selection. UTRs have the highest rate of strongly



**Figure 3.5** Fraction of alleles segregating under different selection regimes by site class in non-overlapping 100 kbp windows for (a, b, c, d) the autosomes and the (e) X chromosome. The selection regimes are strongly deleterious (*d*, in red), weakly deleterious (*b*, in yellow) and neutral (*f*, in green). Blank windows are discarded because no enough data was available.

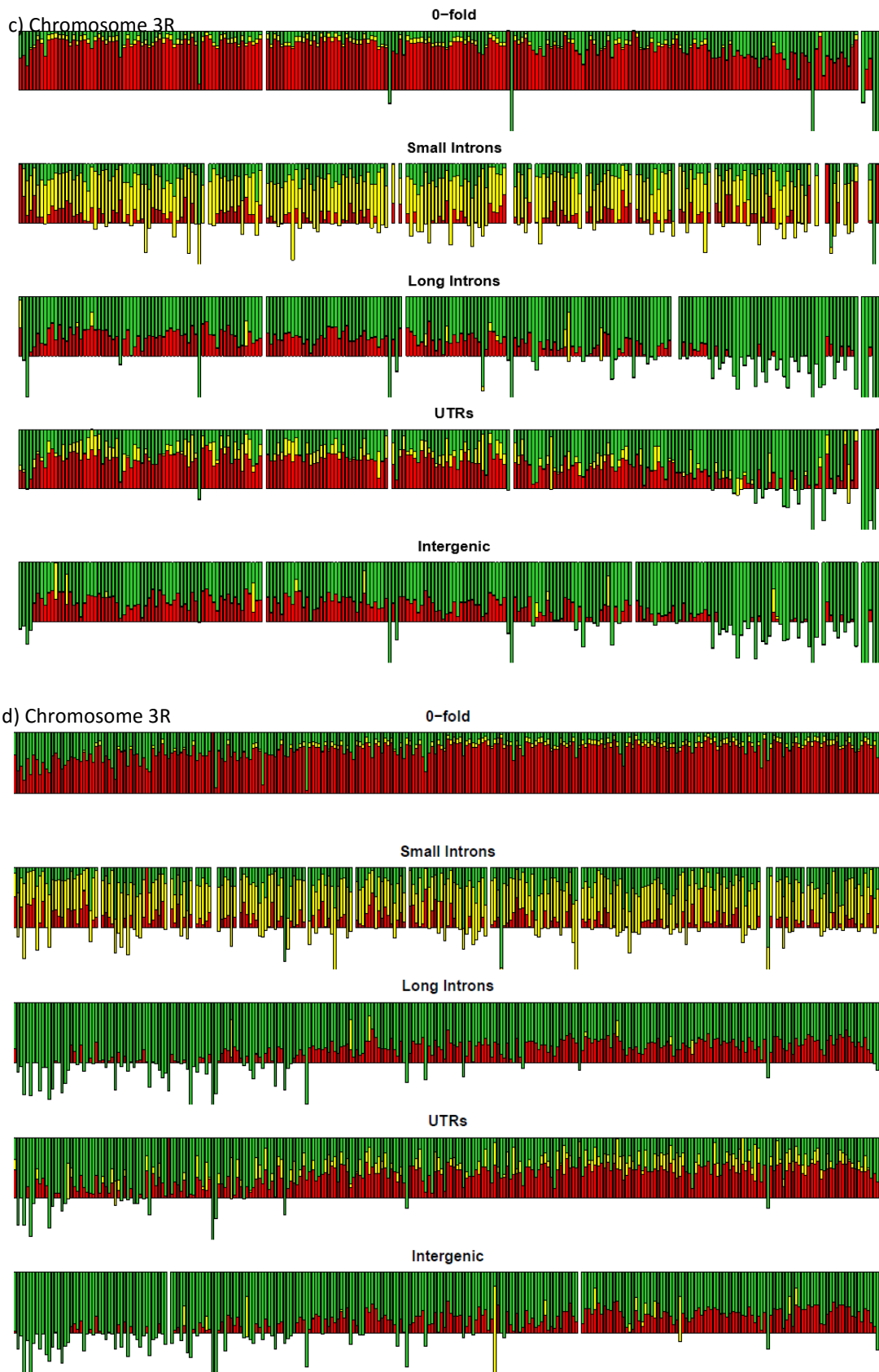


Figure 3.5 (cont)

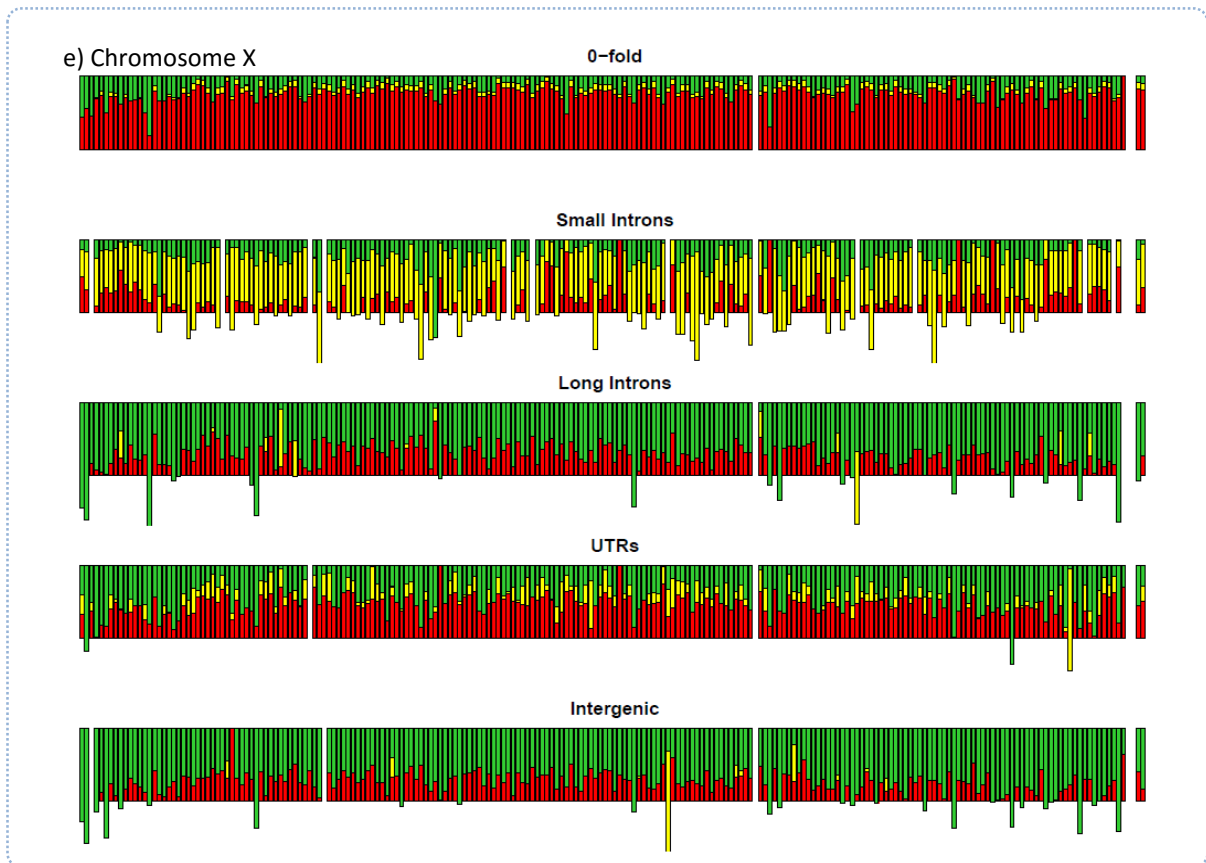


Figure 3.5 (end)

deleterious selection after non-synonymous sites, and the functional gene region with the highest fraction of slightly deleterious alleles, but again their alleles are uniformly distributed along the chromosome. Long introns and intergenic regions behave in the same way; they have a lower fraction of strongly deleterious sites compared to non-synonymous sites and UTRs, and the amount of sites under weak negative selection amounts too little. Finally, short introns display a quite different pattern when compared with the other classes, with an overall excess of weakly deleterious positions (Figure 3.5).

Interestingly, remarkable features of the estimates in figure 3.5 are the regions that have incongruent negative values for some fractions of allele segregating under specific selective regimes. Careful observation shows that this only happens in windows without highly deleterious selection and also seems to be related to an excess of neutral sites since it's also observed in windows without slightly deleterious sites as well. A possible explanation could be that sites in the observed selected class are acting more neutrally than the putatively neutral class. This is more clearly visible with the results of the short intron class, which is suggested to be more neutral than the widely used synonymous sites.

### 3.3 Indel variation landscape in the genome of a natural population of *D. melanogaster*

---

Although being the second most abundant source of genetic variation after SNPs, indels remained understudied due to problems for their reliable detection with current sequencing technologies. Data from DGRP freeze 2 has provided us with 357,608 confident indel variants in 205 genomes from an initial set of 1,296,080 non-SNP variants (see 2.1.1). This indel data set has allowed us to do the most exhaustive genome-wide indel study to date in *D. melanogaster*, and it is comparable with the study of 1.6 million indels from 179 human genomes of Montgomery *et al.* (2013).

#### 3.3.1 Genome-wide indel statistics and distribution

After estimating the ancestral state for 210,268 indels (see 2.4.2) we found that 86% of “deletions” and 74% of “insertions” inferred from the reference genome were true deletions and insertions according to the polarized estimates. Evolutionarily derived deletions ( $n = 145,015$ ; 69%) outnumber insertions ( $n = 65,253$ ; 31%) by 2.2:1 (Table 3.3; Figure 3.6). This estimate of the deletion:insertion ratio for *D. melanogaster* is consistent with previous estimates, which indicates a bias toward higher deletion than insertion rates:

- Petrov (2002) ratio = 8.7:1 from 87 deletions and 10 insertions.
- Ometto *et al.* (2005) ratio = 2.17:1 from 26 intergenic deletions and 12 intergenic insertions and 2:1 from 62 intronic deletions and 31 intronic insertions.
- Assis and Kondrashov (2012) ratio = 3.5:1 from 614 deletions and 179 insertions from gene conversion events;
- Leushkin *et al.* (2013) ratio deletion:insertion = 2.36:1.

There are, on average, 60% fewer deletions ( $\chi_1^2 = 3815, P = 0$ ) and 74% fewer insertions ( $\chi_1^2 = 645.6, P = 0$ ) on the X chromosome than on the major autosomal chromosomal arms (Table 3.3). Although most indels are small (1–2 bp), deletions are, on average, longer than insertions (Table 3.3; Figure 3.6). However, the longest indels are insertions, most of which correspond to P transposable elements which have recently colonized the *D. melanogaster* genome (Kidwell 1993). The longest insertions are preferentially located in centromeric regions.

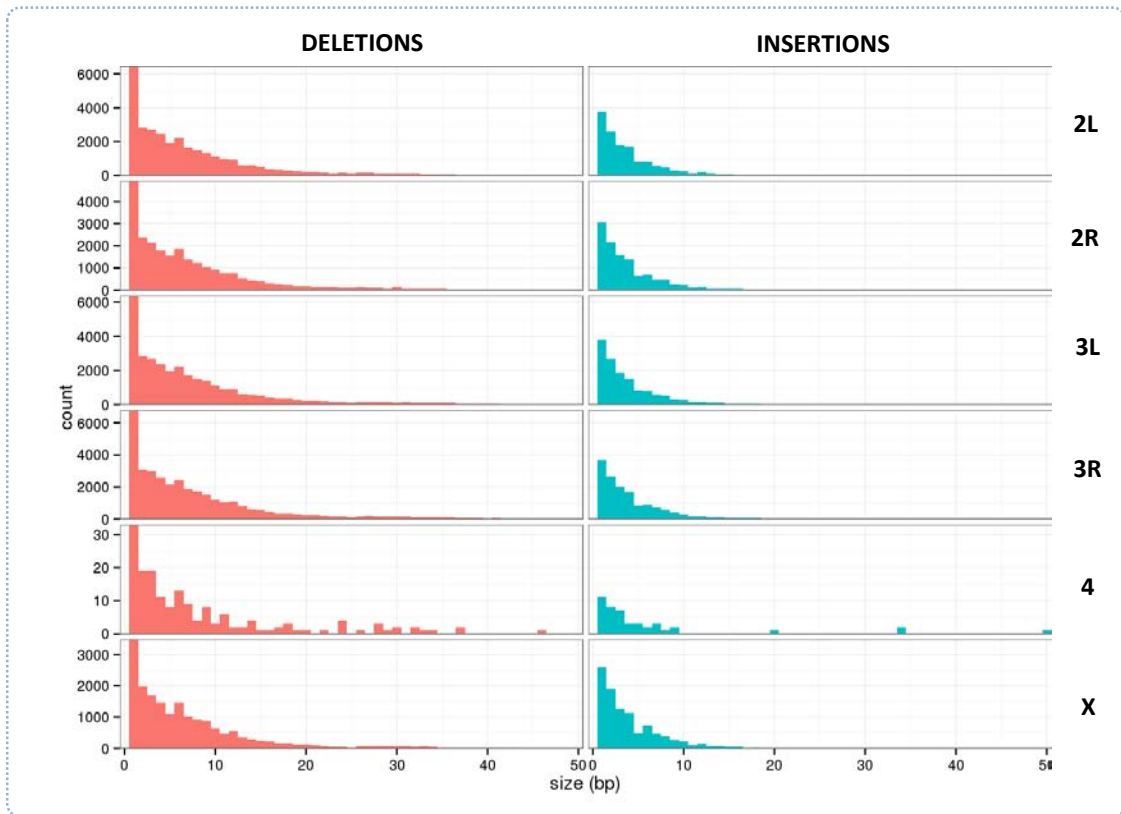
Indel size distributions show the same pattern regardless the chromosome observed (Figure 3.6). The distribution is extremely skewed to the right, with a high number of very small indels



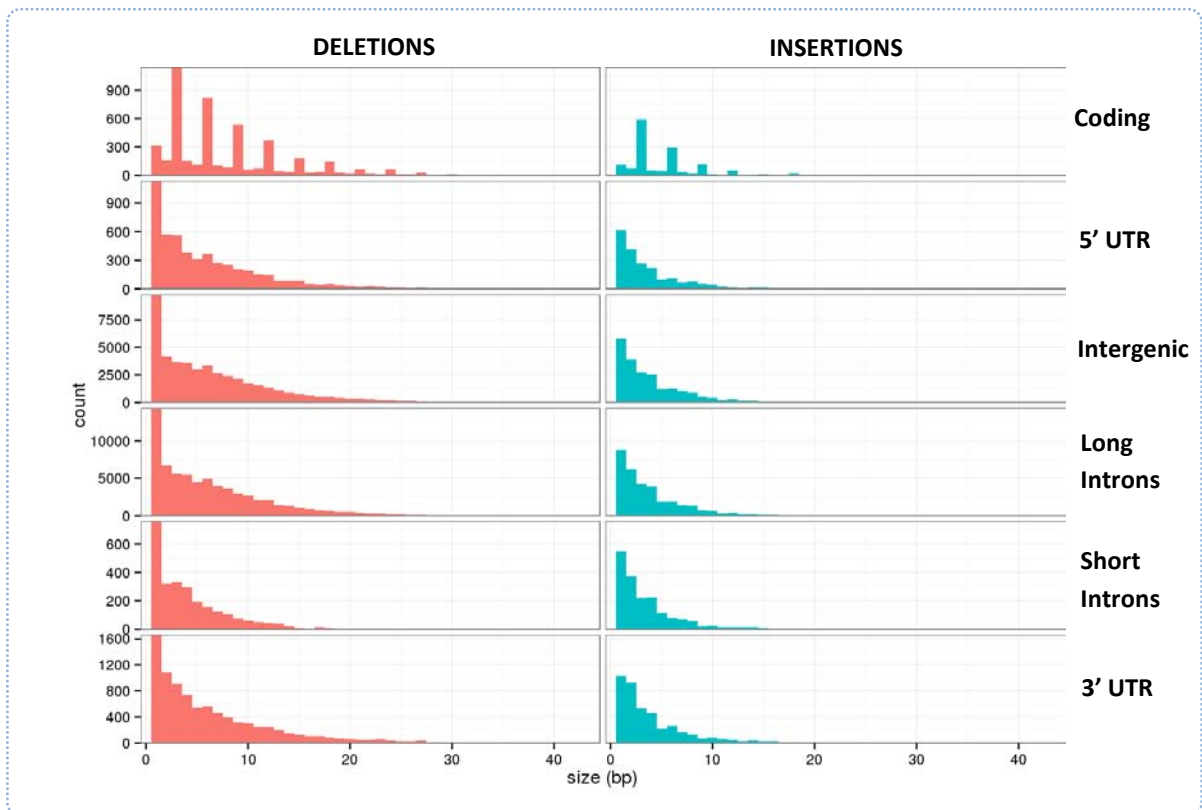
and a rapid decrease in size displayed with a long tail of longer indels. The excess of deletions over insertions can also be observed qualitatively in figure 3.6, especially with the smallest ones of 1 base pair. The size distributions have the same pattern when looking at indels by functional class instead (Figure 3.7). Indels in 3' and 5' UTRs, long and short introns, and intergenic regions, show the observed excess of small indels and a long tail of larger indels. However, a very distinctive pattern appears when looking at indels in coding regions. The size distribution of indels in coding regions has discrete “peaks” for indel sizes in multiples of 3 bp (Figure 3.7). This pattern suggests strong negative selection against coding non-multiple of three nucleotides indels (most of them assumed to cause coding sequence frame-shifts) compared to more relaxed selection for insertions and deletions multiple of three (more probable to span complete codons), a phenomenon previously reported also on DGRP lines (Massouras *et al.* 2012, Leushkin *et al.* 2013) and in humans (Montgomery *et al.* 2013).

**Table 3.3** Indel count and size statistics. Indels are polarized with respect to their ancestral state using *D. simulans* as outgroup.

Chromosome	Total number	Maximum size (bp)	Mean size (bp)	Median size (bp)	Standard deviation (bp)
<b>ALL</b>	210,268	2,921	8.47	4	49.54
<b>Deletions</b>	145,015	2,915	8.93	5	18.34
<b>Insertions</b>	65,253	2,921	7.47	3	84.62
<b>2L</b>	46,225				
<b>Deletions</b>	32,138	459	8.70	5	13.30
<b>Insertions</b>	14,087	2,915	6.45	3	66.59
<b>2R</b>	37,479				
<b>Deletions</b>	25,660	2,843	8.96	6	23.02
<b>Insertions</b>	11,819	2,915	7.17	3	78.92
<b>3L</b>	46,749				
<b>Deletions</b>	32,521	2,915	9.15	5	23.67
<b>Insertions</b>	14,228	2,921	7.35	3	82.77
<b>3R</b>	50,600				
<b>Deletions</b>	35,755	616	9.19	6	14.13
<b>Insertions</b>	14,845	2,915	7.41	3	80.77
<b>4</b>	219				
<b>Deletions</b>	171	89	9.10	5	12.32
<b>Insertions</b>	48	2,915	133.54	3	586.10
<b>X</b>	28,996				
<b>Deletions</b>	18,770	783	8.35	5	14.79
<b>Insertions</b>	10,226	2,915	8.88	3	108.27



**Figure 3.6** Indel size distribution per chromosome arm. Size distribution of insertions and deletions are given per chromosome arm. The bin size is 1 bp. Indels longer than 50bp are not shown.



**Figure 3.7** Indel size distribution by functional class. Size distribution of insertions (IN) and deletions (DEL) per functional class. The bin size is 1 bp. Indels longer than 50 bp are not shown.

### 3.3.2 Covariation between SNP and indel variation

The SNP genotype calls are highly correlated between DGRP Freeze 1.0 (Mackay *et al.* 2012) and Freeze 2.0. Spearman rank order correlations ( $\rho$ ) for estimates of SNP nucleotide polymorphisms ( $\pi$ ) (Nei 1987) among 100-kb non-overlapping windows range from  $\rho = 0.94$  for the X chromosome to  $\rho = 0.99$  for 3R (Table 3.4), therefore, makes sense to use only the Freeze 2.0 nucleotide polymorphism in our estimates from now on. Indel diversity is directly correlated with SNP polymorphism (Table 3.4) and shows the previously observed pattern of higher levels of polymorphism along autosomic chromosome arms and then decreasing gradually as approaching the centromere (Figure 3.8a). Again, this does not seem to be true for the X chromosome where indel polymorphism seems to be uniform only to decrease abruptly in telomere and centromere. Analysing insertions and deletions separately we can observe the same variation pattern along the chromosome arms with high correlation especially in autosomes (Figure 3.8b and c, Table 3.4). All indel estimates in autosomes are also correlated with recombination (Table 3.4), being this correlation weaker in the X chromosome.

**Table 3.4** Spearman correlation matrix among pairs of variables. Data pairs values are estimates for 100 kbp non-overlapping windows along each chromosome arm (X, 2L, 2R, 3L, 3R).  $\pi_{SNP1}$  and  $\pi_{SNP2}$  refer to SNP diversity measures in DGRP Freeze 1.0 and Freeze 2.0, respectively. *Indel* refers to insertions and deletions combined, *In* to insertions only and *Del* to deletions only. All *In* and *Del* calls are for DGRP Freeze 2.0, and polarized with respect to *D. simulans*. *c* is recombination in  $cM Mb^{-1}$ . Entries in the table are Spearman's correlation coefficients (p-value). n/a: not applicable.

Correlation	Chromosome Arm					
	X	2L	2R	3L	3R	4
$\pi_{SNP1}, \pi_{SNP2}$	0.941 (2.2e-16)	0.967 (2.2e-16)	0.976 (2.2e-16)	0.982 (2.2e-16)	0.987 (2.2e-16)	n/a
$\pi_{SNP2}, \pi_{indel}$	0.660 (2.2e-16)	0.836 (2.2e-16)	0.860 (2.2e-16)	0.886 (2.2e-16)	0.880 (2.2e-16)	0.731 (6.32e-3)
$\pi_{SNP2}, \pi_{in}$	0.581 (2.2e-16)	0.824 (2.2e-16)	0.842 (2.2e-16)	0.867 (2.2e-16)	0.868 (2.2e-16)	0.687 (1.20e-2)
$\pi_{SNP2}, \pi_{del}$	0.616 (2.2e-16)	0.756 (2.2e-16)	0.787 (2.2e-16)	0.817 (2.2e-16)	0.795 (2.2e-16)	0.505 (8.12e-2)
$\pi_{in}, \pi_{del}$	0.613 (2.2e-16)	0.751 (2.2e-16)	0.804 (2.2e-16)	0.798 (2.2e-16)	0.758 (2.2e-16)	0.308 (3.06e-1)
$\pi_{SNP2}, c$	0.387 (2.02e-09)	0.608 (2.2e-16)	0.693 (2.2e-16)	0.707 (2.2e-16)	0.644 (2.2e-16)	n/a
$\pi_{indel}, c$	0.383 (2.94e-09)	0.504 (3.15e-16)	0.689 (2.2e-16)	0.727 (2.2e-16)	0.571 (2.2e-16)	n/a
$\pi_{in}, c$	0.376 (6.46e-09)	0.519 (2.2e-16)	0.705 (2.2e-16)	0.748 (2.2e-16)	0.608 (2.2e-16)	n/a
$\pi_{del}, c$	0.334 (2.99e-07)	0.447 (1.11e-12)	0.619 (2.2e-16)	0.649 (2.2e-16)	0.485 (2.2e-16)	n/a

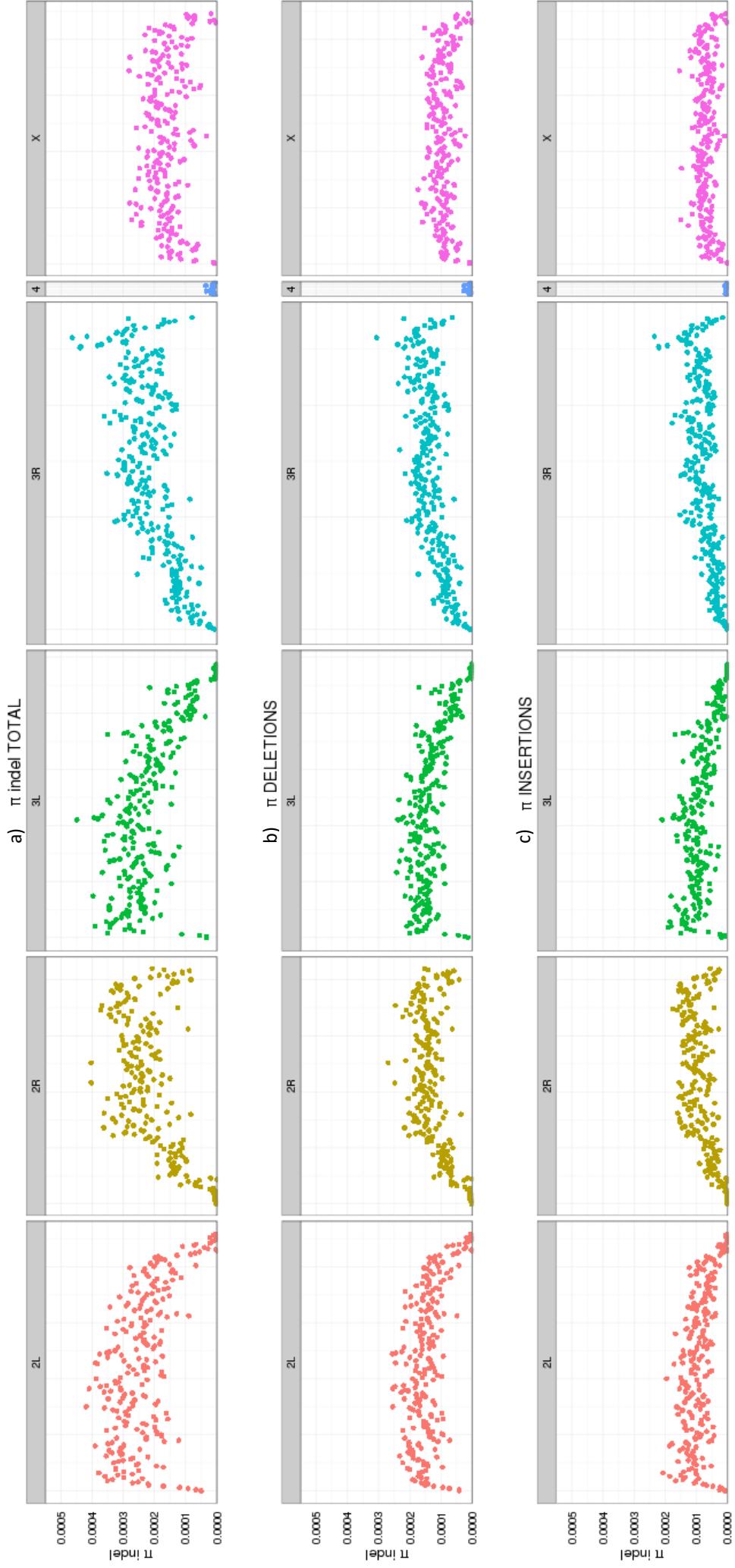
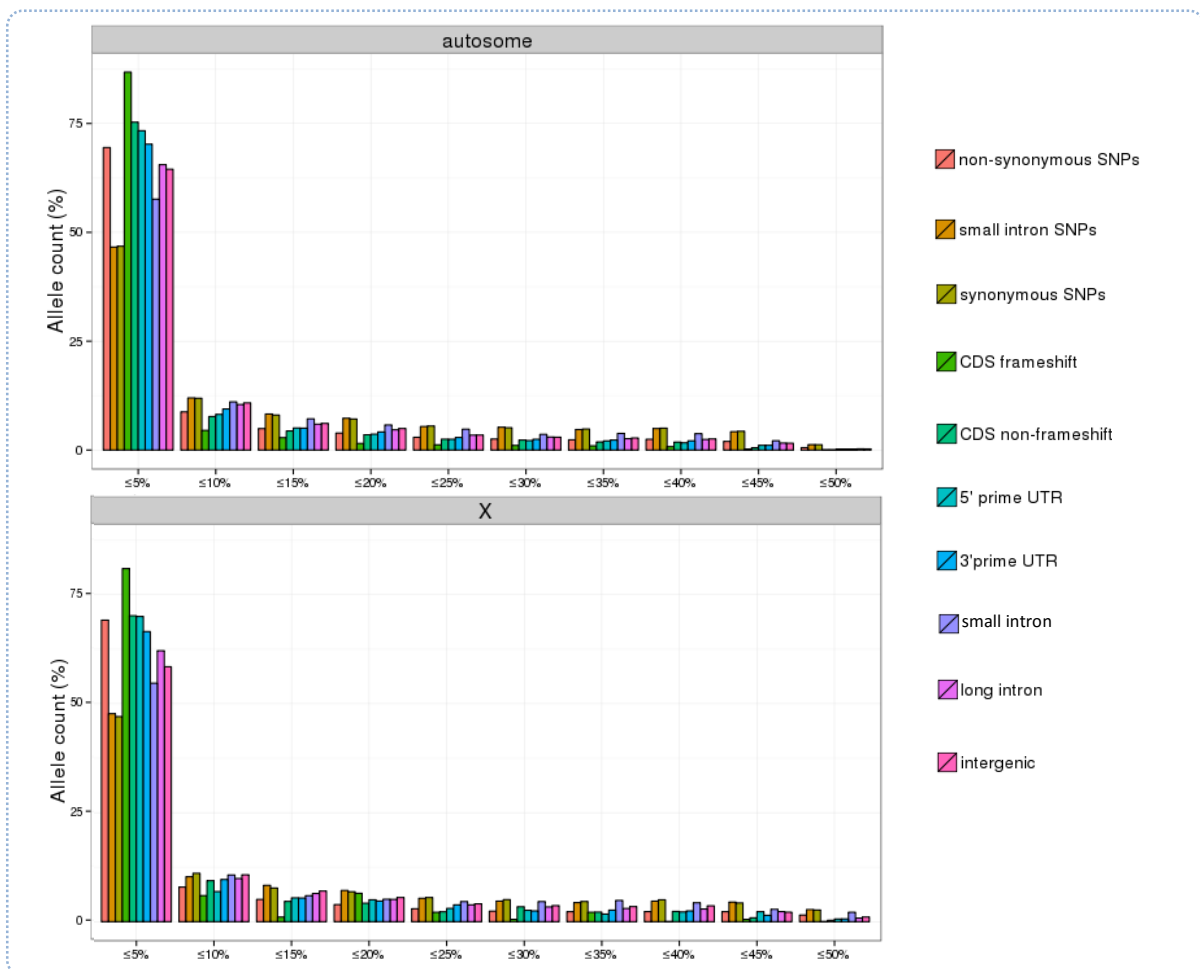


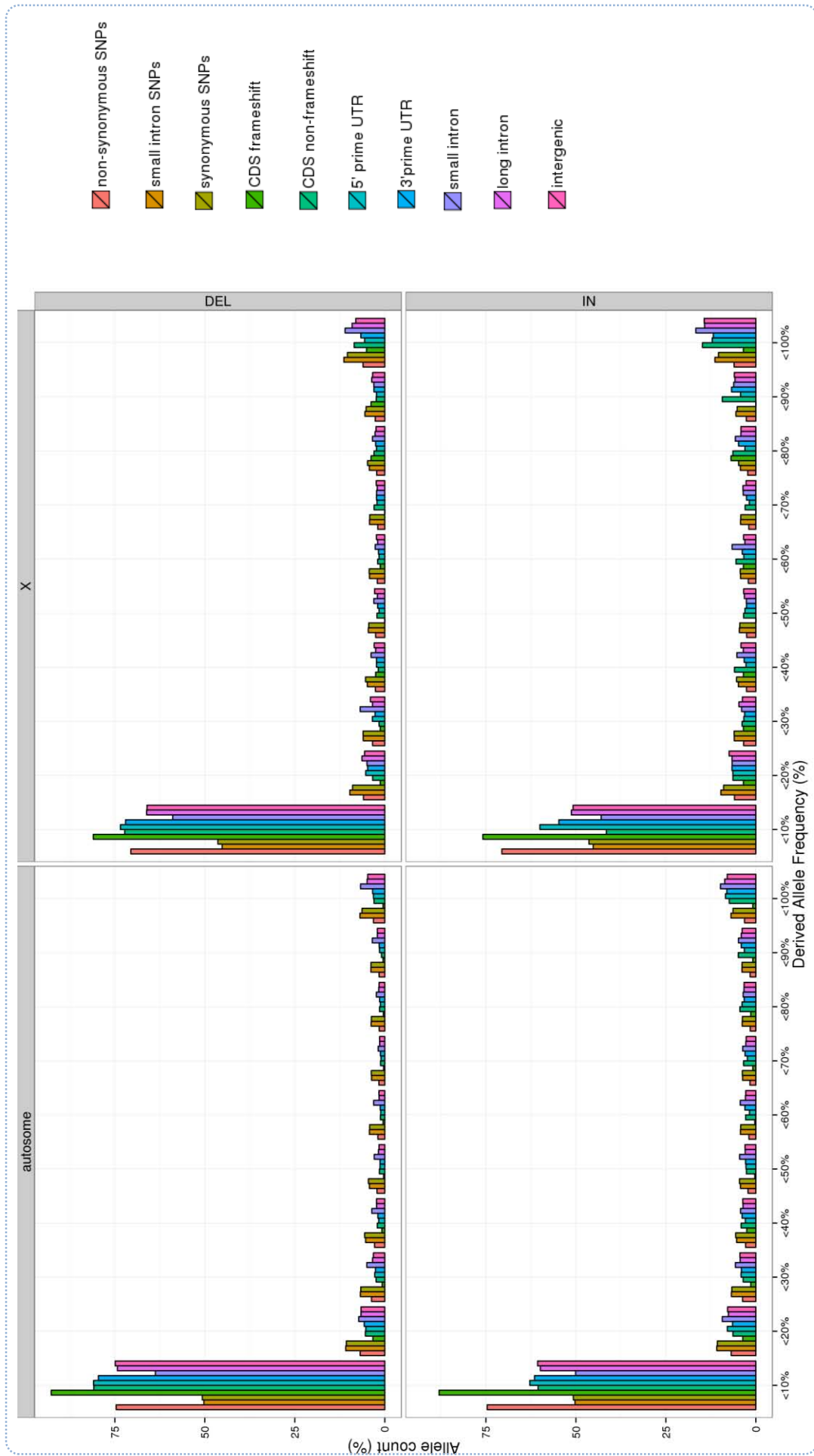
Figure 3.8 Indel diversity ( $\pi_{indel}$ ) values estimated in 100 kb non-overlapping windows along chromosome arms.

### 3.3.4 Derived allele frequency spectrum

The MAF spectra (Figure 3.9) show an excess of low MAF indels compared to SNPs for all functional classes. Given that lower MAF variants are likely enriched for variants under purifying selection, these data are consistent with deleterious fitness effects of indels (Massouras *et al.* 2012). Insertions and deletions with length non multiple of three nucleotides are highly overrepresented among the low DAF class (Figure 3.10), reinforcing the conclusion that negative selection is intense on this indel class. Relative to presumed neutral variants (synonymous SNPs and SNPs in short introns), all deletion classes have an excess of low-frequency derived alleles on all chromosomes. In contrast, the number of low-frequency derived insertion alleles is similar to or less than presumed neutral SNPs for insertions in short introns and multiple of three coding nucleotide insertions on the X chromosome. There is also a slight excess of high-frequency derived insertions compared to SNPs in all chromosomes and all functional categories except for non-multiple of three coding nucleotides. This could indicate more positive selection on insertions than deletions.



**Figure 3.9** Minor allele frequency (MAF) spectra, by functional class, for all indels on the autosomes and the X chromosome.



**Figure 3.10** Derived allele frequency (DAF) spectra for polarized insertions (IN) and deletions (DEL), separately for the autosomes and X chromosome. SNP categories are the same for insertions and deletions.

### 3.3.5 Adaptive selection on indel variation

To do a selection test on our indel dataset we had to select a putatively neutral class. However, there's no clear neutral class for indels. In order to narrow options before the test, we checked global nucleotide polymorphism ( $\pi_{indel}$ ) and divergence ( $k_{indel}$ ) for indels by chromosome and functional class (Table 3.5). High values in both polymorphism and divergence are to be expected in variants acting neutrally. However, the indel classes with highest polymorphism are not all the same ones when looking at the highest divergence indel classes. Only autosomal deletions in short introns (highest polymorphism) and autosomal deletions in intergenic regions (highest divergence) appear between the five highest in both polymorphism and divergence. All short intron indels (autosomic and X insertions and deletions) appear at the top 5 polymorphic classes. Autosomic indels in UTRs appear to have quite divergence values. The next class in both polymorphism and divergence are deletions in autosomic long introns. Since the results were not conclusive to select a unique putatively neutral indel class, we decided to perform the selection tests with DFE-alpha in triplicate, using indels in intergenic regions, long and short introns as neutral classes.

**Table 3.5** Polymorphism ( $\pi$ ) and divergence for indel class and chromosome, ordered from higher to lower values. Abbreviations: AUT. (Autosomes), DEL (Deletions), IN (Insertion),

Polymorphism ( $\pi$ )				Divergence			
Class	Chr	In/del		Class	Chr	In/del	
Short intron	AUT.	DEL	0.000229753	Intergenic	AUT.	DEL	0.001496293
Short intron	AUT.	IN	0.000193738	5' UTR	AUT.	DEL	0.001040506
Short intron	X	IN	0.000151614	3' UTR	AUT.	DEL	0.000976717
Intergenic	AUT.	DEL	0.000146241	Short intron	AUT.	DEL	0.000965923
Short intron	X	DEL	0.000129388	Long intron	AUT.	DEL	0.000954280
Long intron	AUT.	DEL	0.000112138	Intergenic	AUT.	IN	0.000862761
3' UTR	AUT.	DEL	0.000110009	CDS	AUT.	DEL	0.000859433
5' UTR	AUT.	DEL	0.000108455	Intergenic	X	IN	0.000760885
Intergenic	X	DEL	0.000100441	CDS	X	IN	0.000726285
Intergenic	AUT.	IN	0.000086206	Intergenic	X	DEL	0.000691206
5' UTR	X	DEL	0.000076923	5' UTR	AUT.	IN	0.000685704
Long intron	X	DEL	0.000073840	Short intron	AUT.	IN	0.000649590
3' UTR	X	DEL	0.000071869	3' UTR	AUT.	IN	0.000646464
3' UTR	AUT.	IN	0.000070808	CDS	X	DEL	0.000626537
Intergenic	X	IN	0.000068403	5' UTR	X	IN	0.000618709
Long intron	AUT.	IN	0.000067771	Long intron	AUT.	IN	0.000597146
5' UTR	AUT.	IN	0.000061544	CDS	AUT.	IN	0.000595738
3' UTR	X	IN	0.000059172	Short intron	X	IN	0.000594473
Long intron	X	IN	0.000053701	3' UTR	X	IN	0.000540739
5' UTR	X	IN	0.000046490	5' UTR	X	DEL	0.000510184
CDS	AUT.	DEL	0.000016853	Short intron	X	DEL	0.000497871
CDS	X	DEL	0.000013070	Long intron	X	IN	0.000492733
CDS	X	IN	0.000009311	3' UTR	X	DEL	0.000457592
CDS	AUT.	IN	0.000006690	Long intron	X	DEL	0.000415241

The DFE-alpha results in Table 3.6 show us the proportions of adaptive fixations ( $\alpha$ ) and proportions of adaptive indel fixations relative to neutral fixations ( $\omega_\alpha$ ) in our indel dataset. At first glance, values of  $\alpha$  and  $\omega_\alpha$  using the three different neutral classes differ greatly. Estimations using intergenic indels as neutral show low values of  $\alpha$  and  $\omega_\alpha$ , while on the other hand, the results using short and long introns as neutral class display quite high proportions of adaptive fixations for all indel classes.

However, some patterns are clearly visible between the three tests, regardless of the neutral class used. We can observe that indels in coding regions present the highest proportion of adaptive fixations, while the other functional classes display lower values of  $\alpha$  and  $\omega_\alpha$ . Globally, there's no clear pattern indicating differences between deletions and insertions. However, the lowest value of  $\alpha$  in the three tests always corresponds to deletions in coding autosomic regions. This could indicate both differences between X chromosome and autosomes and differences between deletions and insertions globally in the genome. Previous results in this chapter already suggest higher selection efficiency in the X chromosome when compared with the autosomes, but also can have implications in genome size mechanisms (see Discussion).

Overall, introns seem better candidates to neutral indel regions than intergenic indels. Also, taking into account the higher proportion of fixations and the highest values of polymorphism, short indels could be the best candidate for neutral indel class. Still, short indels could have strong constraints in their size, so with these results are not enough to jump to such conclusions (see Discussion).



**Table 3.6** DFE-alpha proportions ( $\alpha$ ) and relative proportions ( $w_\alpha$ ) of adaptive evolution between insertion and deletion functional classes and chromosomes. Each column corresponds with the class used as putatively neutral for indel evolution.

$\alpha$	<i>Intergenic</i>				<i>Long introns</i>				<i>Short introns</i>			
	Deletions		Insertions		Deletions		Insertions		Deletions		Insertions	
	Aut.	X	Aut.	X	Aut.	X	Aut.	X	Aut.	X	Aut.	X
<b>Coding</b>	-0.87658	0.971147	0.837714	0.914222	0.463579	0.9789	0.915286	0.933808	0.274096	0.992238	0.919863	0.976136
<b>Short introns</b>	0.33633	-1.00637	-0.1634	-0.3208	0.279694	0.094592	0.269351	0.261255	-	-	-	-
<b>Long introns</b>	0.415913	0.47183	-0.29223	-0.09889	-	-	-	-	0.616064	0.830143	0.767129	0.669207
<b>3' UTR</b>	-0.71907	0.521752	-0.17035	0.258391	-0.01293	0.651567	0.265198	0.167128	0.714726	0.847301	0.852961	0.77496
<b>5' UTR</b>	0.912763	0.627912	-0.08865	0.081221	0.92899	0.728889	0.314145	0.297088	0.961919	0.883024	0.864275	0.761438
<b>Intergenic</b>	-	-	-	-	0.706637	0.718634	0.472668	0.452804	0.842134	0.878388	0.632042	0.706472
$w_\alpha$	<i>Intergenic</i>				<i>Long introns</i>				<i>Short introns</i>			
	Deletions		Insertions		Deletions		Insertions		Deletions		Insertions	
	Aut.	X	Aut.	X	Aut.	X	Aut.	X	Aut.	X	Aut.	X
<b>Coding</b>	-0.46712	1.118777	0.786419	1.103821	0.427354	2.058649	1.301866	1.881522	0.243187	3.07135	1.711951	1.755824
<b>Short introns</b>	0.216849	-0.50159	-0.14045	-0.24288	0.282545	0.104475	0.368646	0.353648	-	-	-	-
<b>Long introns</b>	0.267543	0.29457	-0.22615	-0.06265	-	-	-	-	0.740559	1.295472	1.063323	0.625278
<b>3' UTR</b>	-0.41829	0.365627	-0.14555	0.190388	-0.01276	0.833584	0.36091	0.200665	0.902664	1.495053	1.378108	0.847554
<b>5' UTR</b>	0.682329	0.510757	-0.08143	0.068116	1.111883	1.082396	0.458033	0.422654	1.386253	1.836715	1.506884	0.955665
<b>Intergenic</b>	-	-	-	-	1.176338	1.438117	0.896339	0.827499	1.6876	2.457297	1.252029	1.056365



---

## **Part 4.**

### **Discussion**

---

## 4. Discussion

The DGRP genome dataset is a formidable resource for population genomics. In a natural population of model species *D. melanogaster* we have described the patterns of polymorphism and divergence (nucleotide and indels) along chromosome arms, the relationships between diversity and recombination, mapped the footprint of natural selection on SNP and non-SNP variants throughout the genome and developed a complete, public and accessible map of the polymorphism present in this population.

### 4.1 Population genomics software development

---

#### 4.1.1 PopDrowser and Genome Brower

A new dimension to genetic variation studies is provided by the new availability of within-species complete genome sequences. Next-generation sequencing technologies are making affordable genome-wide population genetics data, not only for humans and the main model organisms, but also for most organisms on which research is actively carried out on genetics, ecology or evolution (Pool *et al.* 2010).

The *Drosophila* Genetic Reference Panel (DGRP) (Mackay *et al.* 2012) has sequenced and analyzed the patterns of genome variation in 158 (freeze 1) and 205 (freeze 2, see section 2.1.3 of Material and methods) inbred lines of *Drosophila melanogaster* from a single population of Raleigh (USA), and conducted a genome-wide association analysis of some phenotypic traits. A major goal of this project is to create a resource of common genetic polymorphism data to aid further population genomics analyses. As a part of this DGRP project, we have implemented a modified GBrowse, a generic genome browser interface, specifically designed for the automatic estimation and representation of population genetic variation in *D. melanogaster*.

A number of web-based genome browsers displaying genetic variation data are already available (Benson *et al.* 2002, Kent *et al.* 2002, Hubbard *et al.* 2002, Stein *et al.* 2002, Frazer *et al.* 2007, Dubchak and Ryaboy 2006). Such browsers, however, are not well suited to deal with population genomics sequence information. For example, HapMap (The International HapMap Consortium 2003), the most comprehensive genome browser of variation data so far, contains information on SNPs, CNVs, and linkage disequilibrium of human populations. It does not offer, however, genetic variation estimates along sliding-windows or neutrality-based tests.

Unlike other population analysis tools (Hutter *et al.* 2006, Kofler *et al.* 2011), the PopDrowser is a genome browser specially designed for the representation and analysis of population genomics data. Originally, it comprised 180 tracks including the most commonly-used summary statistics and tests of the population genetics theory (see section 2.1.3 of Material and Methods) and standard and integrative MK test for the genome-wide detection of natural selection at any region. It is especially appropriate either for genome-wide analyses (*e.g.* mapping natural selection, detecting regions undergoing positive selection or functional unannotated regions that are highly constrained, or correlating levels of variation with other genomic measures, etc.), or for the detailed analysis on small chromosome regions. The flexibility of administrating the visualization of tracks, adding custom tracks of data, or analyzing specific regions on-the-fly, allows accommodating the PopDrowser to the user's needs and facilitating their analyses.

The technology of the GBrowse has been proved useful to graphically display huge quantities of population genetics information in an easy and interactive fashion. An advantage of the Gbrowse technology is its open source nature. This allowed us modify parts of the code to customize some types of visualization, and has even allowed us the creation of a new functionality not present by default in the browser: the ability to perform on-the-fly analysis and the download of fragments from the complete DGRP freeze 1.0 alignment.

One disadvantage, not exclusive of the GBrowse but shared between all the genome browsing systems, is the lack of standardized solutions, consequence of this continuous 'reinventing the wheel' common with this kind of tools and mentioned in the introduction. We could observe an example of it when the problem of displaying high volumes of data appeared during these years that we developed PopDrowser. During the first years in the already short history of genome browsing, browsers handled relatively low quantities of data since displaying some annotations in its local genomic context were not a computational problem (some gene names and start and end coordinates and that's it). Also, the visualization of really large regions is counterproductive

because too many annotations visualized at the same time are generally too small and has no value for the user. Even the growing number of tracks was no problem in this regard, at least, not until high throughput data arrived. New NGS technologies provided us with huge volumes of data, even for small regions with not many annotations. The lack of a standardized format to store, access and visualize such high volumes of data makes difficult to work with it. In the case of the GBrowse, in the last four years they have changed three times the standard storing format for big volumes of data. This is one of the most important challenges that genome visualizations face right now, and some promising solutions have already been proposed like the BigWig data format (see Box 3, in Chapter 2: Materials and Methods) or the use of high performance computing and parallelization algorithms to optimize the access to the data like in the relatively recent GenomeMaps browser (Medina *et al.* 2013).

Another caveat is trying to maintain the base GBrowse code constantly up-to date. This is an intrinsically consequence of the initial design of the tool (a problem also shared with many Linux/Unix based non-binary tools). Once configured, a genome browser is a very flexible, powerful and easy to use tool: adding new information and tracks is really straight forward, and it can be done without affecting the experience of the users. However, trying to upgrade the whole base code is no easy task and, in fact, very time consuming, since best experience is only assured in static environments (operation system-wise speaking). Since GBrowse is based in a collection multiple of bioperl scripts, custom changes to the code (at first a great feature to be available to do) are lost after an upgrade of the system. Also, GBrowse developers could introduce changes that affect our configurations, so an active administrator has to tackle this issues after an upgrade as well. An example of this can be observed with the PopDrowser: described containing 180, in the actual public version dozens are disabled since a change in the visualization had conflicts with negative values (affecting most of the tracks in the neutrality tests section). To correct this, both a complete upgrade of the GBrowse system and a conversion of all the precomputed estimates are required. Time constraints and the need to work on other parts of the thesis lead to this PhD. candidate to hold that upgrade, and probably leave it to a next member of the lab.

Another issue is the portability of these systems. By definition, as web interfaces, the actual software resides in a single server computer, and users simply request the information to this server from their personal terminals. The increase in computational power of computers and the fairly low resources requirements of a GBrowse installation makes the option of the

installation into a virtual machine a feasible solution to allow the sharing of the complete implementation to people interested in use and modify it for their own uses.

Probably the most relevant next improvements in genome browser technologies will come more from the side of the software/hardware than the biological side, especially in visualization and user interaction with the data. On the visualization side, it's clear that the current system based on the reference genome is insufficient:

1. Genome Browse systems are originally designed to display single genome information while more and more individual genome data is produced.
2. Genome Browsers are not well designed to display annotations with nucleotides not present in the reference, like new sequence of insertions or copy number variation.
3. The only visualization of every genome browser is based in single region navigation. Interactions between different regions or chromosomes within the species, or even between species are not referred to.

Clever use of different glyphs can patch the problem for now, but maybe these tricks will quickly become obsolete. Another huge change we can expect as well is the arrival of information on genome 3D organization, which for sure will force some changes in the visualization of genomic data that to date is mainly based in 2D.

#### 4.1.2 Integrative-MKT implementation

Our first java approach to implement the integrative-MKT is useful by showing the fraction of alleles segregating under different selection regimes, but some inconsistencies in the estimates indicate that the results must be interpreted with caution (see section 2.3.3 of Material and Methods, section 3.2.3 of Results and Figure 3.5). This conclusion comes from the values of selection that span outside their graphic area. Interestingly, these 'out of range' estimates only happen in regions where the presence of highly deleterious alleles could not be inferred, suggesting that this is the effect of some biological mechanism underneath. A possible explanation could be that sites in the observed selected class are acting more neutrally than the putatively neutral class. In our case, the putatively neutral class are 4-fold degenerated coding sites. It could be possible that in coding regions, some 4-fold degenerated sites may be in linkage with non-synonymous strongly selected sites, thus overestimating the neutrality of these sites.

This effect could be revealed when these putatively neutral sites are compared with sites acting even more neutrally. This could explain the qualitative differences observed within the short intron sites, these sites could effectively be “more neutral than the neutral” class of choice. Mutations in 4-fold sites, without functional effect, have been considered selectively neutral since long (Kimura 1968, King & Jukes 1969), hence their use as neutral variation in many studies like in Mackay *et al.* (2012). However, codon usage bias observations suggested that synonymous mutations could have some function (Hershberg & Petrov 2009). Recently, Lawrie *et al.* (2013) estimated that 22% of mutations in 4-fold positions in *D. melanogaster* are deleterious enough to disappear rapidly from the genome. That 4-fold positions could have some functional impact is another explanation to the observation that a given class could be more neutral than the putatively neutral selected class.

Another aspect to consider comes from comments on the Integrative-MKT method from Campos *et al.* (2014). They suggested that the assumptions behind the our method lead to an important underestimation of the fraction of deleterious sites, since in integrative-MKT deleterious sites are assumed to be always removed from the population, but there’s a possibility that a proportion of deleterious sites could remain segregating. However, since this underestimation is constant in our calculations, it does not change our overall conclusions. Clearly, a calibration of the neutrality of the selected class and the sub estimation the strongly deleterious sites could improve the performance of the estimates.



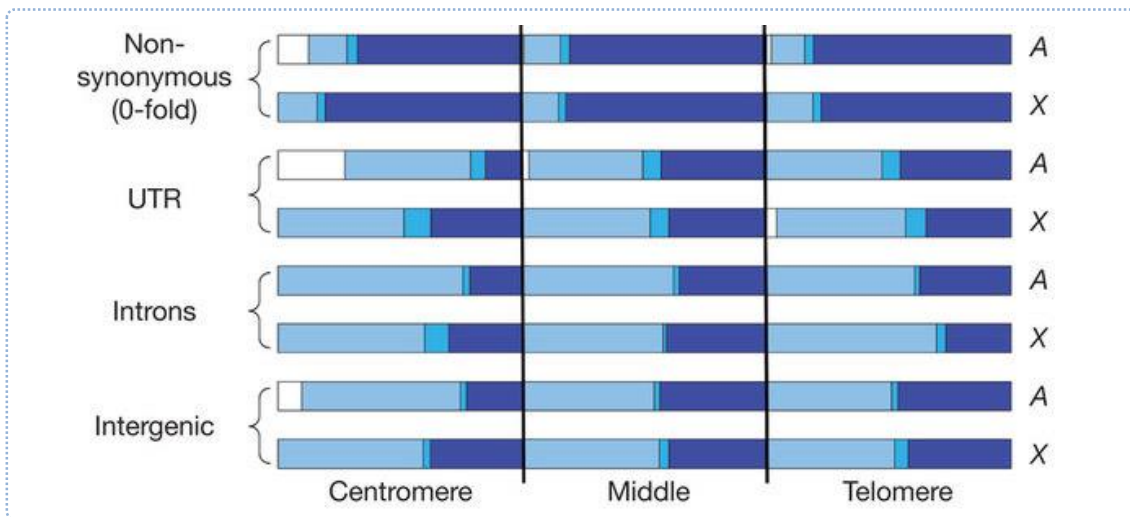
### 4.2.1 Genome-wide patterns of polymorphism and divergence and recombination

From our nucleotide variation analysis we can observe clear different patterns of polymorphism and divergence between (i) centromeric vs. non-centromeric regions within autosome arms, and (ii) between autosomes vs. X chromosome (Figure 3.4, Table 3.1). In autosomes the polymorphism levels are maintained across chromosome arms, but then decrease as reaching the centromeric and telomeric regions. The reduction is really abrupt in the telomere, in contrast with the reduction towards the centromere: it's more gradual and it does not affect only the centromere but approximately a third of the chromosome arm is affected. Chromosome X has an overall reduced level of polymorphism, also maintained across the chromosome. It also exhibits an abrupt reduction of polymorphism in the telomere and somewhat in the centromere as well, not showing the gradual reduction of polymorphism observed in the autosomes. Interestingly, we observe a fairly similar pattern for recombination in autosomes (sudden increase in the telomeres, elevated through the chromosome arm, and gradual decrease towards the centromere) (Figure 3.4a).

This high resolution map of polymorphism is clearly in line with the predicted pattern of reduced polymorphism in the centromeric areas made with only 17 observations by Hudson and Kaplan (1995), at least in the autosomes (Figure 1.5). This polymorphism pattern is correlated with recombination levels in the autosomes (Table 3.2) and, as exposed in the introduction, follows more than two decades of observations of reduced polymorphism in areas of low recombination such as the centromeres (See Introduction 1.1.4; Aguade *et al.* 1989; Stephan and Langley 1989; Berry *et al.* 1991; Begun and Aquadro 1992; Martin-Campos *et al.* 1992; Stephan and Mitchell 1992; Langley *et al.* 1993). The correlation polymorphism-recombination is not so clear in the X chromosome (Table 3.2, see discussion for the X chromosome differences in section 4.2.3). Looking in detail at our correlations between polymorphism and recombination, it seems to exist a recombination threshold around  $2 \text{ cM Mb}^{-1}$  above which  $\pi$  does not increase further, reaching the amount of polymorphism a plateau. Below this threshold, nucleotide diversity correlates strongly with recombination, attaining its lowest values in zero-recombining regions. Our study strongly confirms at a genome-wide scale the observations of other studies that recurrently find correlation between polymorphism and recombination in the *Drosophila* genus (Begun & Aquadro 1992, Begun *et al.* 2007, Kulathinal *et al.* 2008, Sackton *et al.* 2009, Stevison & Noor 2010). This reduction in polymorphism correlated with the recombination is not limited to the

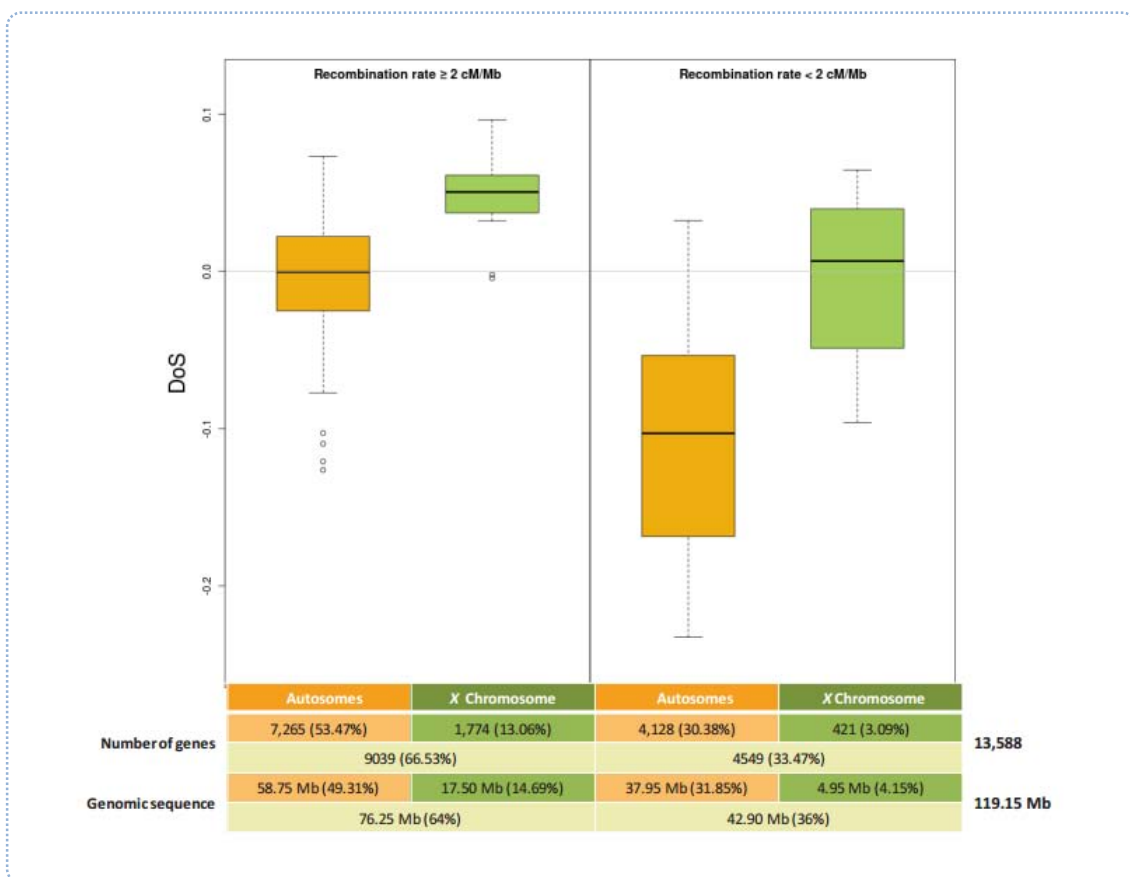
regions with zero recombination or to the centromeres. Instead, the correlation spans at least a third of the autosomes (Barron 2015). It was already suggested that recombination could be the major force shaping the polymorphism patterns in the genome (Sackton *et al.* 2009). However, analyses with the DGRP data confirms and reinforce this hypothesis, since it has been observed that recombination explains x3 times (in autosomes) and x8 times (in the X chromosome) more the amount of polymorphic explained by other factors (Barrón 2015).

As mentioned in the introduction (See section 1.1.5), when an adaptive allele is fixed, or a deleterious allele removed from the population, if there is no recombination, neutral nearby variation is reduced. The lower the recombination, the longer the region affected and more pronounced the reduction of polymorphism. Furthermore, if linked regions are long enough, there is the possibility that both events of adaptive and deleterious selection can occur together, interfering with each other and affecting as well the linked neutral variation, an effect called Hill-Robertson Interference (HRI) and producing a reduction of the efficiency of selection (see 4.2.2). Mackay *et al.* (2012) and Castellano *et al.* (in press) have shown that those regions whose recombination rate is under the threshold are associated with lower efficiency of selection, especially in autosomes, as shown in figures 3.6 and 3.7 (Figure 4.1 and 4.2).



**Figure 4.1** The fraction of alleles segregating under different selection regimes by site class and chromosome region, for the autosomes (A) and the X chromosome (X). The selection regimes are strongly deleterious (*d*, dark blue), weakly deleterious (*b*, blue), recently neutral ( $\gamma$ , white) and old neutral ( $f - \gamma$ , light blue). Each chromosome arm has been divided in three regions of equal size (in Mb): centromere, middle and telomere [from Mackay *et al.* 2012 Figure 3].

Looking deeper at the polymorphism-recombination correlation, Barrón (2015) observed that the threshold on which the correlation vanishes is different in each chromosome (ranging from  $3.36 \text{ cM Mb}^{-1}$  in chromosome arm 2L to  $1.54 \text{ cM Mb}^{-1}$  in chromosome X), and also that the reduced efficiency of selection is not exclusive of the centromeric areas. As a consequence, when analysing both variation data and recombination rate along the genome, it can be divided in one hand, (i) regions with high recombination and low linked selection (thus, with high efficiency of selection), and on the other hand (ii) regions with low recombination and increased linked selection events with lower efficiency of selection (see 4.2.2).



**Figure 4.2** Test for differences in the Direction of Selection (DoS) between autosomes and the X. DoS was calculated for each independent gene and averaged every 50 gene windows along chromosome arms. Values are given for the X chromosome and autosomes, for regions of high ( $\geq 2 \text{ cM Mb}^{-1}$ ) and low ( $< 2 \text{ cM Mb}^{-1}$ ) regions of recombination [from McKay *et al.* 2012, sup. figure 8].

Finally, divergence shows no variation along the chromosomes except for a peak of divergence in the centromeric region that could be explained by an increased mutation rate, and/or more sequencing errors and/or or higher fixation of slightly deleterious mutations due to the reduction of the efficiency of selection in low recombination region (Birky & Walsh 1988). No correlation is found between recombination and divergence (Table 3.2), and, looking at the precomputed data in the PopDrowser, there is actually more sequencing errors in the centromeric regions (Figure 3.4b, Figure 3.1 and “Ns” track in the PopDrowser). So we could assume that those peaks in the divergence around the centromere are at least in part due to artifactual variants.

#### 4.2.2 Mapping natural selection across the genome and the major effect of recombination

Our genome-wide selection estimates using the integrative-MKT method suggest the presence of adaptive fixations, neutral variation and deleterious alleles all across the genome (Figure 3.4c, Figure 3.5). Selection seems to be pervasive all along the genome, but its strength and mode can vary broadly when considering the different functional classes of sites. In correspondence with our window-by-window observations, gene-by-gene approaches (Figure 3.6, Mackay *et al.* 2012) show, for example, that the most constrained site class are the non-synonymous sites with 73.9% of sites being strongly deleterious ( $d$ ), while in non-coding sites  $d$  ranges from 38.1% in intergenic regions to 31.8% in introns. We observe significant shifting on the importance of the different selective fractions when comparing centromeric and non-centromeric regions of the autosomes. In centromeric regions, and regardless of the site class considered, the fraction of strongly deleterious sites is reduced considerably (Figure 3.6). This reduction is remarkable in UTR sites and is still important in non-synonymous sites. The diminution of  $d$  is compensated by an increase in the fraction of neutral or nearly neutral sites ( $f$ ).

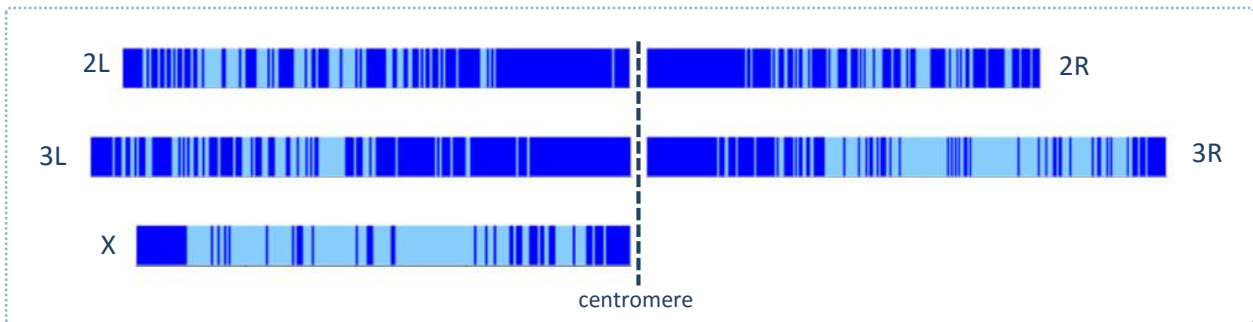
The integrative-MKT method allows quantifying the average proportion of sites that have ‘recently’ become neutral ( $\gamma$ ), from the time of separation of *D. melanogaster* and *D. yakuba*, and that have resulted in an excess of polymorphism relative to divergence, distinguishing these sites from those that have undergone slightly deleterious mutations. In Mackay *et al.* (2012) (Figure 3.6), a higher fraction of sites that have recently become neutral or nearly neutral is observed in centromeric regions (27% in UTR sites and 13% in non-synonymous sites when compared to non-centromeric mutations 1.5% in UTR sites and 1% in non-synonymous sites). Even though the low or null recombination in these regions is expected to decrease the efficiency of natural selection and to account for the higher percentage of neutral or nearly

neutral sites, this “recently neutral” excess is expected to be reflected in divergence as well. However, we find an excess of polymorphism relative to divergence. This seems to indicate that weakly deleterious selection is common in centromeric regions. We also infer that an intense redistribution of the regimes of selection is taking place, converting sites that were under strongly deleterious selection and weak selection into neutral or nearly neutral sites in recent times. For adaptive selection, in Mackay *et al.* (2012), it is estimated that 24% of fixations ( $\alpha$ ) in coding regions (looking at 50kb windows) are adaptive.

Assuming that the mutation rate does not change with recombination (from the observed patterns of divergence, Table 3.2 and figure 3.4b), it has been suggested that the reduction of variation in low recombination regions is consequence of the effect of linked selection mechanisms such as hitchhiking and background selection (Begun & Aquadro 1992, Charlesworth *et al.* 1993). In such regions, the low or lack of recombination makes selection very inefficient due to the H<sub>Ri</sub>. Accordingly, two or more selected variants which are in linkage disequilibrium interfere with each other reducing the efficiency of natural selection and as a consequence variation is reduced in those segments of the genome (Hill & Robertson 1966; see section 1.1.5 of Introduction). As recombination increases, the linkage disequilibrium between alleles is reduced, variants can segregate more freely as we can observe in the non-centromeric regions of the autosome arms and in the X chromosome, and selection can act more efficiently. As stated before, a reduction of the efficiency of selection in low recombination areas could lead to an increased fixation of slightly deleterious variants simply by drift. Once a variable in a low recombination region is fixed, variation from other segregating variants is reduced in a longer linked region, thus the reduction of variation we observe in the centromeric regions of the autosomes in *Drosophila*.

As observed by Barrón (2015), the genome can be divided into two distinctive types of regions with opposite molecular evolutionary dynamics: linked selection blocks (LSB) and non-linked selection blocks (NLSB) (Figure 4.3). In NLSB, the classical interpretation of genetic variation based on the neutral theory as a null model can be applied (Cavalli-Sforza 1966; Lewontin and Krakauer 1973), while in LSB H<sub>Ri</sub> is predicted to occur recurrently. Around 40% of the *D. melanogaster* genome seems to be constituted by NLSB (27% in autosomes, 77% in the X chromosome). Hence, 60% of the genome, especially in the autosomes, seems to be in a sub-optimal situation regarding natural selection efficiency. This implies that it is not correct to consider the nucleotide as the unit of selection (Bustamante *et al.* 2001; Hahn 2008; Neher 2013; Messer & Petrov 2013) and that linked selection should be taken into account while trying to

infer natural selection. Furthermore, recently Castellano *et al.* (in press) have estimated, for the first time, the overall impact of H<sub>Ri</sub> on the efficiency of selection in the whole genome of *D. melanogaster*. Looking at the rate of adaptive evolution ( $\alpha$ ), they calculated that H<sub>Ri</sub> diminishes the rate of adaptive evolution by ~24%, and that this fraction depends on the gene mutation rate: genes with low mutation rates lose ~17% of their adaptive substitutions while genes with high mutation rates lose ~60%.

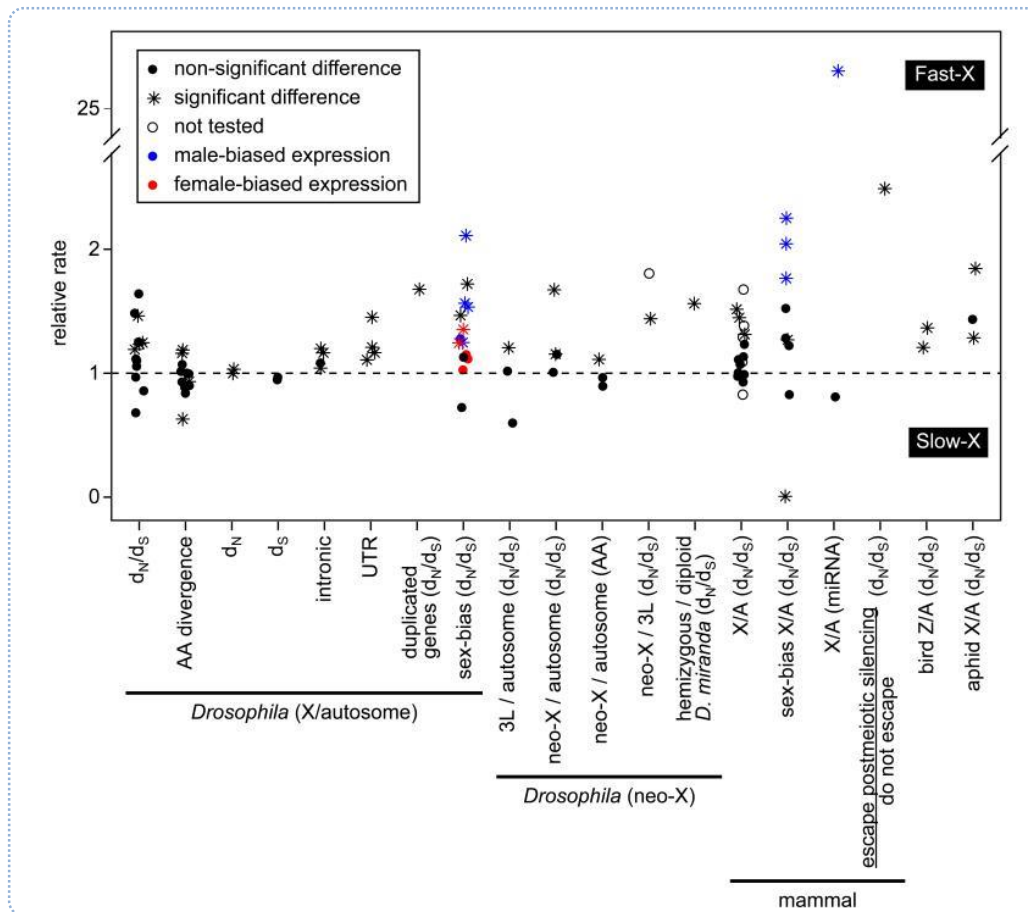


**Figure 4.3** Linked selection blocks (LSB, in deep blue color) map of the genome of *D. melanogaster*, based in the recombination rate ( $c$ ) in windows of 100Kb [from Barrón 2015, figure 3.10]

#### 4.2.3 Fast-X hypothesis

The X chromosome exhibits a completely different variation pattern compared with the autosomes ones. As mentioned before, the overall polymorphism is reduced when compared with the autosomes, and there's no gradual reduction in polymorphism towards the centromere. Also, the X chromosome has a higher average rate of recombination ( $2.9 \text{ cM Mb}^{-1}$  over the  $2.1 \text{ cM Mb}^{-1}$  of the autosomes) and divergence ( $k_{\text{dyak}} 0.1403$  over  $k_{\text{dyak}} 0.1256$  of the autosomes). Moreover, the X chromosome has less neutral sites and a corresponding higher percentage of strongly deleterious alleles in general (Figure 3.6, Mackay *et al.* 2012). These observations could support the 'Fast-X' hypothesis. This hypothesis appears from the observation that, although usually similar to the autosomes in size and cytogenetic appearance, the hemizygosity of the X chromosome in males expose new partially recessive mutation to a larger intensity of selection that may cause increased rates of evolution (Charlesworth *et al.* 1987, Vicoso & Charlesworth 2006, Meisel & Connallon 2013). In other words, the 'Fast-X' hypothesis proposes that X-linked genes can be more divergent between species when compared with autosomal genes. However, several comparisons of autosome and

X divergence rates in *Drosophila* and other species have led to contradictory results (Thornton *et al.* 2006, Baines *et al.* 2008, Meisel & Connallon 2013) (Figure 4.4). Our results and most recent studies with genomic data in *Drosophila* support the ‘Fast-X’ hypothesis, especially when looking at male-expressed genes. However, the faster X evolution does not seem to be so clear in mammals and birds (Meisel & Connallon 2013).



**Figure 4.4** Tests for faster-X evolution. The relative rate of evolution is plotted for different classes of nucleotide sites and chromosomes in *Drosophila*, mammals and birds. The rate of evolution is measured as either  $d_N/d_S$ , amino acid (AA) divergence, or nucleotide divergence at different classes of sites (indicated on the x-axis). The expectation if X-linked genes and autosomal genes evolve equally is represented by the dashed line [from Meisel and Connallon 2013, figure 2].

Our results show that in *Drosophila melanogaster* the X chromosome exhibits lower polymorphism levels except in centromeric region, and larger molecular evolution than autosomes, which clarifies the previous inconclusive studies based on partial data where chromosome arm regions could not be monitored in detail. This different pattern of polymorphism cannot be explained by mutation rate changes along the chromosome. Although

the X chromosome has higher divergence rate than the autosomes, divergence remains constant across the chromosome arm. Besides, there are no evidence of differences between X and autosomes in mutation accumulation experiments (Keightley *et al.* 2009; Schrider *et al.* 2013). The X chromosome contains a higher percentage of gene regions undergoing both strongly deleterious and adaptive evolution, and a lower level of weak negative selection and relaxation of selection than autosomes in all the arm regions.

In terms of linked and non-linked selection blocks, Barrón (2015) shows that the larger rate of recombination in the X with respect to the autosomes makes that 77% of sites in the X chromosome are free of linked selection events. That is, 50% more sites are selectively independent in this chromosome than in the autosomes. Higher recombination rates imply increased efficiency of selection due to less linked selection and reduced H<sub>Ri</sub>, as also suggested in Mackay *et al.* (2012) and other studies (Langley *et al.* 2012, Pool *et al.* 2012), supporting the hypothesis of a faster evolution of the X chromosome. It must be taken into account, though, that there's only one copy of the X chromosome in males. This implies that, technically, the effective size of the X chromosome in the population is  $\frac{3}{4}$  than of the autosomes. As said previously, in males new mutations are directly exposed to the effect of selection, contributing to different patterns of evolution in the X chromosome (Campos *et al.* 2014, Vicoso & Charlesworth 2006). The increased selection on partially recessive alleles in hemizygotic males together with the higher efficiency of selection due to the larger recombination rate in the X chromosome compared with the autosomes, can act synergically and account for the faster X evolution.



### 4.3 Indel Variation Landscape in *D. melanogaster*

---

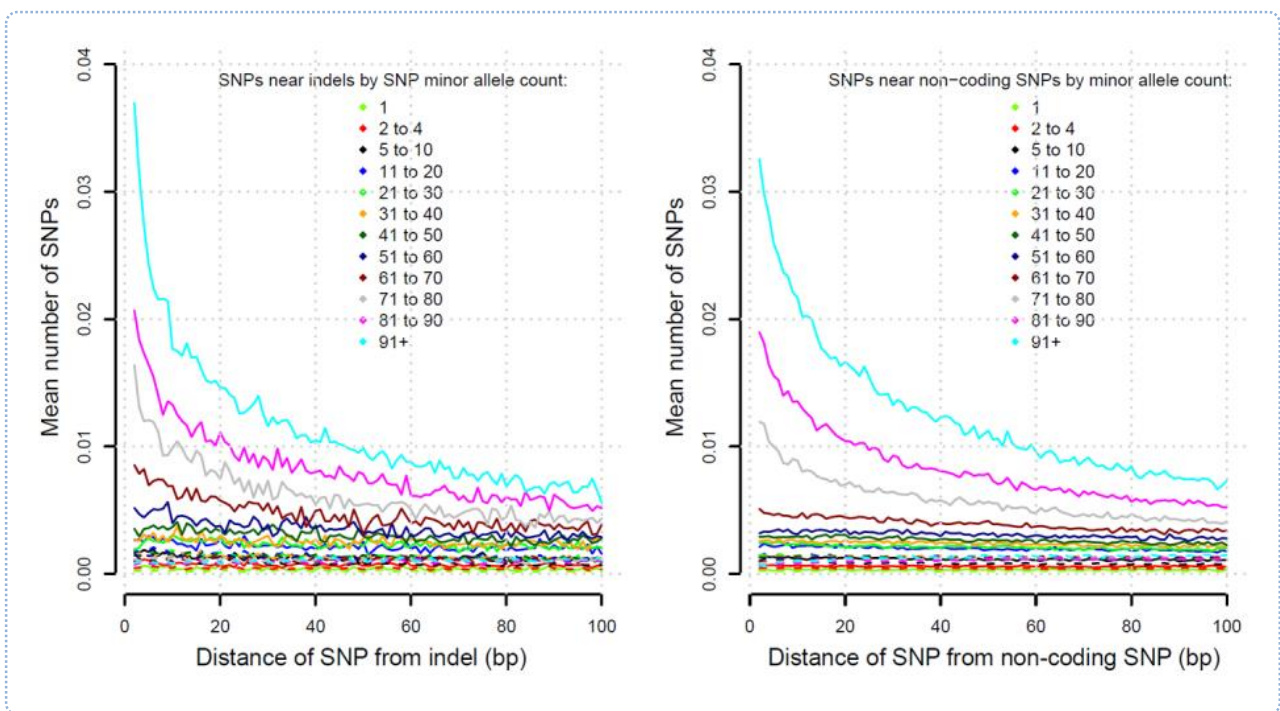
#### 4.3.1 Genome-wide description of indel variation and indel-SNP relationships in *D. melanogaster*

One of the first remarkable results from our indel analysis is the observed excess of deletions vs insertions in the genome. The strong observed bias toward deletions does not seem to be not an artifact due to larger difficulty of calling big insertions than deletions. Respect the reference sequence, around the same numbers of insertions and deletions were called in average for each DGRP Freeze 2.0 line (31682 deletions and 31704 insertions, in average per line), except for the largest variants (>400bp), where more deletions than insertions relative to the reference were called (85 deletions and 98 insertions, in average per line, Huang *et al.* 2014). Thus, a calling bias only exists for large indels, since such variants are a very small fraction (1.72%  $\geq$  100bp non-SNP variants of the DGRP Freeze 2 dataset), this cannot account for the excess of evolutionarily derived deletions. This excess of 2.2:1 deletions over insertions is consistent with previous studies in smaller data sets that also observed higher deletion than insertion rates (Petrov 2002, Ometto *et al.* 2005, Assis and Kondrashov 2012, Leushkin *et al.* 2013; see section 3.3.1 of Results). Again, we observe differences between autosomes and the X chromosome.

In the Freeze 1.0 analysis, we found that SNP nucleotide polymorphism ( $\pi$ ) was reduced near centromeres and telomeres and was positively associated with local recombination rate (for recombination rates  $< 2 \text{ cM Mb}^{-1}$ ) (Figure 3.4a, Table 3.2). The pattern of  $\pi_{indel}$  along chromosomes is similar to that of SNP nucleotide diversity (Figure 3.10). There is a strong positive correlation between indel and nucleotide diversity for all chromosome arms (Table 3.4), even when looking at insertions or deletions separately. Evolutionary models of hitchhiking and background selection predict a positive correlation between recombination and polymorphism for all variants, either indel or SNPs (Begun and Aquadro 1992; Charlesworth *et al.* 1993). We confirmed our previous observation of Freeze 1.0 (Mackay *et al.* 2012) that SNP polymorphism is positively correlated with the local recombination rate, and also extended this observation to insertions and deletions (Table 3.4).

Thus, local recombination rate affects the same way of all types of variants, suggesting that the same evolutionary processes affecting SNPs are the most likely explanation for the observed clustering of indel variants (See section 4.2.1 and 4.2.2). The lack of correlation between recombination and divergence for SNPs and indels (Spearman  $r = 0.037$  genome-wide,  $P = 0.205$ ) excludes a mutagenic effect of recombination to explain such patterns.

Levels of indel and SNP polymorphisms are correlated along the chromosome arms (Table 3.4), that is, SNPs and indels high and low variation regions are mutually clustered. It appears that the clustering of SNP and indel variation is ubiquitous in prokaryotes and eukaryotes (Tian *et al.* 2008, Hodgkinson and Eyre-Walker 2011, McDonald *et al.* 2011, Jovelin and Cutter 2013). Various mechanisms have been proposed to explain this: (i) indels may be mutagenic, either because they induce errors during the DNA polymerase replication near them (Tian *et al.* 2008; Jovelin and Cutter 2013, McDonald *et al.* 2011, Yang and Woodgate 2007) or may cause additional point mutations when segregating due to pairing problems during meiosis (Hodgkinson and Eyre-Walker 2011), (ii) it might be that the regions in which SNPs and indels occur are inherently mutagenic (Hodgkinson and Eyre-Walker 2011), or (iii) SNPs and indels variation patterns behave in parallel because are driven by the same population genomic processes, such the common local recombination rate (Hodgkinson and Eyre-Walker 2011).



**Figure 4.5** Clustering of SNPs nearby selected indels and SNPs. Average number of SNPs at a distance to a high frequency variant (40-50% MAF). Solid lines represent SNP counts in lines that have the variant and dashed lines the SNP counts in lines that do not have it [from Huang *et al.* 2014, Figure 3].

Massouras *et al.* (2012) and Huang *et al.* (2014) have tested with DGRP freeze 2.0 data the hypothesis whether indels are or not mutagenic. They tested if there was an excess of SNPs near selected indels at high frequencies (supposed to be older) when compared to the proportion of

other SNPs near selected SNPs, also at high frequencies. It is expected an increase in low frequency (more recent) variants if there is any increase of mutation rate. No clear differences were found, high allele frequency SNPs cluster around other variants at high frequency (SNPs or indels) and low frequency variants also cluster together, suggesting that indels are not especially more mutagenic for point mutations than SNPs are as well, indicating that variant clustering is not driven by indels. Looking at the same regions in lines without the selected alleles, they observe fewer variants of any frequency than in the same regions with the allele present, refuting the hypothesis that both indels and SNPs occur in regions with increased mutation rate. Then, both types of variants seem to be exposed to the same population genomic processes (Massouras *et al.* 2012, Huang *et al.* 2014) (Figure 4.5).

#### 4.3.2 Selection acting on indels

The site frequency spectrum distributions of indels can give us hints about the selective mechanisms responsible for their variation patterns. In the DAF spectrum we observe an excess of rare alleles in deletions when compared to insertions (Figures 3.11 and 3.12). These results suggest that natural selection acts differently on insertions and deletions, with stronger purifying selection on deletions. Similar patterns have found by Petrov (2002), Assis and Kondrashov (2012) and Leushkin *et al.* (2013) in smaller data sets. This is consistent with the mutational equilibrium theory for genome size evolution proposed by Petrov (2002). According to this theory, deletions are, in general, expected to be more deleterious. Small deletions or insertions in coding regions are almost assured to disrupt genes, and large deletions would affect genes more frequently than small insertions (Petrov 2002). The longer the deletion, the more probability to eliminate functional DNA, hence the more deleterious it is. On the other hand, insertions would have the same probability to affect a coding region regardless of their size, since they only have one breakpoint. Longer insertions may be favoured since the expansion of a genomic region (like in low recombination areas) has more chance to have its LD reduced. At the end, steady equilibrium is reached when the higher mutation rate of deletion over insertion is counterbalanced by a stronger selection against DNA loss compared with a more relaxed selection for DNA gains (Petrov 2002). Flow cytometry analysis has also shown huge variation in total genome sizes with a skew towards larger genomes when compared with the reference strain (Huang *et al.* 2014). This could be an evidence for the presence of weaker selection acting on longer insertions. However, we must be cautious with this result since flow cytometry

measures both euchromatic and heterochromatic DNA, and heterochromatic DNA is rich in transposable elements activity which could explain the variation in genome size.

After seeing the effects of slightly deleterious indels, we wanted to estimate the amount of adaptive fixations of indels using the DFE-alpha software. However, the problem when trying to apply a method like the DFEalpha (which, like the MKT, requires two classes of sites, one being putatively neutral) is to assign the putative neutral indel class. That's not a problem with SNPs, where it's commonly assumed that the most neutral positions are either polymorphism in synonymous coding sites or SNPs in the positions 8-30 of introns  $\leq 65\text{bp}$  (Parsch *et al.* 2010) in a region. There is no a priori clear neutral indel class identified yet. In a first thought, the equivalent of synonymous SNP sites in indels are, apparently, indels in coding regions that does not shift the reading frame. The presence of segregating non-frameshift indels indicates that some proteins are flexible in adding or losing a few amino acids and maintain function, probably because these amino acids are not in important functional sites of the protein (Figure 3.9). However, where a synonymous SNP implies no change in an amino acid, this type of indels imply the gain or loss of one or multiple entire codons, and the neutrality of such loss/gains may be arguable (Zhao *et al.* 2013, Bermejo-Das-Neves *et al.* 2014, Boschiero *et al.* 2015). Also, frame and non-frame shifting indels must be used with caution, because their correct classification is really sensible to errors in annotation of exons and genes (Zhao *et al.* 2013, Boschiero *et al.* 2015).

Finally, for our DFE-alpha analysis we have chosen as the more putatively neutral indels those spanning both short and long introns and intergenic regions. These three classes are the ones with highest divergence/polymorphism ratio (if we do not take into account the high divergence rate in autosomic deletions in UTRs, Table 3.5), and they show the most similar behaviour to the putatively neutral SNP classes when we look at the DAF analysis (Figure 3.12).

Regardless of the neutral class used, the DFEalpha estimates of  $\alpha$  and  $\omega_\alpha$  show that the class category with highest proportion of adaptive fixations are the indels in coding sequence (Figure 4.6). Since indels are expected to be more deleterious, any indel segregating at medium or high frequency in a coding region has high probability to be advantageous, hence it will become fixed quickly. We can also observe a highest level of adaptive fixations in deletions when compared with insertions, probably reflecting as well that deletions undergo more often both positive and negative selection pressure over insertions.

However, despite indels in coding regions exhibit the highest rate of adaptive fixations, also they have the lowest estimated values of  $\alpha$  and  $\omega_\alpha$  within a coding indel class: deletions in autosomal coding regions. This low proportion of adaptive indel fixations in autosomic coding regions could indicate a higher proportion of slightly deleterious deletions in the region (also suggested by the MAF and DAF analyses, Figures 3.11 and 3.12). Also, this lower proportion of adaptive fixations observed for deletions, but not for insertions, could be another argument for the hypothesis that deletions are more selected against than insertions. This could be in favour of the hypothesized equilibrium between selection (against deletions) and more permissiveness to insertions, and this could be a major force maintaining genome size, at least in *Drosophila*.

At the end, the answer to which indel class is more neutral remains unclear. From the three used classes, long intron indels show the most equilibrate results in  $\alpha$  and  $\omega_\alpha$  estimates, but results with short introns have similar patterns, only with higher values. Also, both insertions and deletions in short introns have the highest rates of polymorphism (Table 3.5), suggesting more neutrality than long intron indels. Still, it must be taken into account that short introns may be under size constrain (Comeron and Kreitman 2000) and any indel there is unlikely to act neutrally. However, the effect in short indel size might not be relevant in our study since most of the indels used are of small size. Finally, intergenic regions are probably too heterogeneous in size and composition (when compared to introns that are more delimited), hence the disparity in  $\alpha$  and  $\omega_\alpha$  estimates, however, it's not clear that indels within intergenic regions are less neutral than within introns. These results are only the starting point to the genome-wide analysis of indel variation. Further studies will follow to broaden our understanding of the evolution of indel variants in the genomes.



---

## **Part 5.**

# **Conclusions**

---

## 5. Conclusions

### PopDrowser

1. We have implemented PopDrowser, a genome browser based on GBrowse that has been specially designed for the representation and analysis of population genomics data. PopDrowser automates the estimation of several genetic variation measures along each chromosome from a set of aligned intraspecific sequences and the aligned sequence of outgroup species. The DGRP genome data and the genome sequences of *Drosophila yakuba* and *Drosophila simulans* have been used as the source of polymorphic and outgroup genomics data, respectively.
2. PopDrowser allows the administration and visualization of multiple tracks in an easy and flexible way. A powerful and innovative function of this browser is that it allows performing analyses on-the-fly at any region with user defined parameters. The capabilities of visualization of annotations, track integration and on-the-fly analyses make PopDrowser a useful tool to gain a better comprehension of the population genomic processes at different genome scales.

### Nucleotide variation analysis of DGRP Freeze 1.0

3. We observe a clear and consistent pattern of genome nucleotide diversity ( $\pi$ ) along arms of the autosomic chromosomes: nucleotide diversity is reduced on average 2.4-fold in centromeric regions relative to non-centromeric regions, and at the telomeres. This pattern is not observed in the X chromosome, where diversity is almost uniform along it. Divergence is rather uniform along all chromosome arms.
4. There is a correlation between polymorphism and recombination along chromosome arms. However, it seems to be a threshold around  $2\text{cM Mb}^{-1}$  above which the correlation polymorphism/recombination vanishes. Recombination rate seems to be the major force shaping the patterns of polymorphism along chromosome arms and its effect seems to be mediated by the size of blocks of linked selection.
5. Natural selection, both adaptive and purifying selection, is pervasive in the genome of *Drosophila melanogaster*. Selection is more efficient in the X chromosome as a whole, and in the central and telomeric regions of the autosomes than in the centromeres.



6. Indel size frequency distributions are similar for each functional class of sites except in coding regions, where discrete 'peaks' of indels whose size are multiple of three are observed. This distinctive indel size pattern in coding regions suggests a strong negative selection against frame shifting indels compared with a more relaxed selection for insertions and deletions spanning complete codons.
7. The parallel levels of SNPs and indel diversity along chromosome arms seem to obey a common underlying population genomics factor, being recombination rate this main factor.
8. A strict protocol to infer whether an indel variant is a derived deletion or derived insertion has been implemented. Our estimates show that deletions outnumber insertions according to a ratio deletion-to-insertion 2.2:1 in all chromosomes. These results strongly confirm previous studies suggesting higher deletion rates in the genome of *D. melanogaster*.
9. Natural selection acts differently between insertions and deletions, being deletions more strongly selected by purifying selection. This is consistent with the mutational equilibrium theory for genome size evolution, which proposes that optimal genome size is maintained by the trade-off between purifying selection acting on small deletions exhibiting higher mutation rate and looser selection acting on insertions appearing in lower rates.



---

## Bibliography

- Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 21: 974–984 (2011)
- Achaz G. Frequency spectrum neutrality tests, one for all and all for one. *Genetics* 183, 249–258 (2009)
- Aguade M, Miyashita N and Langley CH. Reduced Variation in the Yellow-Achaete-Scute Region in Natural Populations of *Drosophila melanogaster*. *Genetics*, 122(3): 607–615 (1989)
- Akey JM, Zhang G, Zhang K, Jin L and Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12, 1805–1814 (2002)
- Alkan C, Coe BP and Eichler EE. Genome structural variation discovery and genotyping. *Nature Reviews* 12, 363–375 (2011)
- Altschul FS, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J mol boil* 215(3): 430–410 (1990).
- Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ and Donnelly P. A haplotype map of the human genome. *Nature* 437: 1299–1320 (2005)
- Andolfatto P. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Molecular Biology and Evolution*, 18(3):279–90 (2001)
- Andolfatto P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437, 1149–1152 (2005)
- Andolfatto P and Przeworski M. Regions of lower crossing over harbor more rare variants in African *Drosophila melanogaster*. *Genetics* 158, 657–665 (2001)
- Assis R and Kondrashov AS. A strong deletion bias in nonallelic gene conversion. *PLoS Genet* 8: e1002508. (2012)
- Avery OT, MacLeod M and McCarty M. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a deoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *Journal of Experimental Medicine* 79: 137–158 (1944).
- Avise JC and Lansman RA. Polymorphism of mitochondrial DNA in populations of higher animals. In Nei, M. and Koehn, R. K. (ed.) *Evolution of Gene and Proteins*, Sinauer, Sunderland, Mass., pp. 147–164. (1983)
- Baines JF, Sawyer SA, Hartl DL and Parsch J. Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in *Drosophila*. *Molecular Biology and Evolution*. 25:1639–1650 (2008).
- Bamshad M and Wooding SP. Signatures of natural selection in the human genome. *Nat Rev Genet*, 4(2): 99–111 (2003)
- Barrón MG. Nucleotide variation patterns and linked selection blocks mapping along the *Drosophila melanogaster* genome. *PhD Thesis*, Universitat Autònoma de Barcelona (2015)
- Barton NH, Briggs DEG, Eisen JA, Goldstein DB and Patel NH. *Evolution*. Cold Spring Harbor Laboratory Press (2007)
- Beaumont MA and Nichols RA. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B*, 263, 1619–162 (1996)
- Beaumont MA and Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* 13, 969–980 (2004)
- Begun DJ and Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, 356(6369): 519–520 (1992)
- Begun DJ and Aquadro CF. African and North American populations of *Drosophila*

- melanogaster* are very different at the DNA level. *Nature*, 7 365(6446):548-50 (1993)
- Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, Pachter L, Myers E and Langley CH. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in *Drosophila simulans*. *PLoS Biology* 5(11):e310.doi:10.1371/journal.pbio.0050310 (2007)
- Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology* 5, e310 (2007)
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA and Wheeler DL. GenBank. *Nucleic Acids Research* 30, 17-20 (2002)
- Bermejo-Das-Neves C, Nguyen H-N, Poch O, Thompson JD. A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i). *BMC Bioinformatics*. 15,111 doi:10.1186/1471-2105-15-111 (2014)
- Berry AJ, Ajioka JW and Kreitman M. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* 129(4), 1111-1117 (1991)
- Birky CW and Walsh JB. Effects of linkage on rates of molecular evolution. *PNAS* 85(17), 6414-6418 (1988)
- Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE and The Mouse Genome Database Group. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res* 42(D1):D810-D817 (2014)
- Boschiero C, Gheyas AA, Ralph HK, Eory L, Paton B, Kuo R, Fulton J, Preisinger R, Kaiser P, Burt DW. Detection and characterization of small insertion and deletion genetic variants in modern layer chicken genomes. *BMC genomics* 16, 562 (2015)
- Bustamante CD, Wakeley J, Sawyer S, and Hartl DL. Directional Selection and the SiteFrequency Spectrum. *Genetics* 159 (4): 1779-1788 (2001)
- Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M and Clark AG. Natural selection on protein-coding genes in the human genome. *Nature* 437, 1153–1157 (2005)
- Campos LJ, Halligan DL, Haddrill PR and Charlesworth B. The relation between recombination rate and patterns of molecular evolution in *Drosophila melanogaster*. *Molecular Biology and Evolution* 31, 1010-1028 (2014)
- Casillas S. Development and application of bioinformatic tools for the representation and analysis of genetic diversity. *PhD Thesis*, Universitat Autònoma de Barcelona (2007).
- Casillas S and Barbadilla A. PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. *Nucleic Acids Res.* 34, W632-634 (2006)
- Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A and Eyre-Walker A. Adaptive evolution is substantially impeded by Hill-Robertson interference in *Drosophila*. *Mol Biol Evo* (in press)
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421 (2009)
- Charlesworth B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res*, 63(3), 213-227 (1994)
- Charlesworth B. Molecular population genomics: a short history. *Genetic Research*, 29, 397-411 (2010)
- Charlesworth B, Coyne JA and Barton NH. The relative rates of evolution of sex chromosomes and autosomes. *The American Naturalist* 130(1), 113-146 (1987)
- Charlesworth B, Morgan MT and Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4), 1289-1303 (1993)
- Charlesworth J and Eyre-Walker A. The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol* 25, 1007-1015 (2008)

- Cavalli-Sforza LL. Population structure and human evolution. *Proceedings of the Royal Society of London*, 164: 362-379 (1966)
- Chen H, Patterson N and Reich D. Population Differentiation as a test for selective sweeps. *Genome Research* 20, 393-402 (2010)
- Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S and Wong ED *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40 (Database issue):D700-5 (2012)
- Chin G. Double strand break repair. *J Biol Chem*, 272, 24097-24100 (1997)
- Clark AG et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–218 (2007)
- Collins FS, Brooks LD and Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research* 8(12): 1229-1231 (1998)
- Comeron JM and Kreitman M. The correlation between long intron length and recombination in *Drosophila*: equilibrium between mutational and selective forces. *Genetics*, 156: 1175-1190 (2000)
- Comeron JM., Williford A and Kliman RM. The Hill-Robertson Effect: Evolutionary Consequences of Weak Selection and Linkage in Finite Populations. *Heredity* 100(1), 19–31 (2008)
- Comeron JM, Ratnappan R and Bailin S. The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genet* 8: e1002905. (2012)
- Conrad DF and Hurler ME. The population genetics of structural variation. *Nature Genetics* 39, S30 - S36 (2007)
- Conrad DF, Pinto D, Redon R *et al.* Origins and functional impact of copy number variation in the human genome. *Nature*, 464:704-712 (2010)
- Consortium T. C. E. S. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282(5396), 2012-2018 (1998)
- Craig DW, Pearson JV, Szelinger S, *et al.* Identification of genetic variants using barcoded multiplexed sequencing. *Nature methods* 5(10), 887-893 (2008)
- Darwin C. On the origin of species by the means of natural selection. Or the preservation of favoured races in the struggle for life. John Murray, London (1859).
- David JR and Capy P. Genetic variation of *Drosophila melanogaster* natural populations. *Trends in Genetics* 4(4):106-11 (1988)
- De Velasco B, Shen J, Go S and Hartenstein V. Embryonic development of the *Drosophila* corpus cardiacum, a neuroendocrine gland with Bibliography 215 similarity to the vertebrate pituitary, is controlled by sine oculis and glass. *Dev Biol*, 274(2): 280-294 (2004)
- Dobzhansky T and Sturtevant AH. Inversions in the Chromosomes of *Drosophila Pseudoobscura*. *Genetics* 23(1): 28-64 (1938)
- Dobzhansky T. Genetics and the Origin of Species. *Columbia University Press, New York* (1937)
- Dobzhansky T. Genetics of the Evolutionary Process. *Columbia University Press* (1970)
- Dowell RD, Jokerst RM, Day A, Eddy SR and Stein L. The distributed annotation system. *BMC bioinformatics* 2, 7 (2001)
- Drosophila* 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203-218 (2007)
- Dubchak I and Ryaboy DV: VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. *Methods in molecular biology (Clifton, N.J.)* 338, 69-89 (2006)
- Duchen P, Živković D, Hutter S, Stephan W and Laurent S. Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population. *Genetics*, 193(1): 291–301 (2013)

- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5), 1792-1797. (2004)
- Egea R, Casillas S and Barbadilla A: Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic acids research* 36, W157-162. (2008)
- Ellegren H. Genome Sequencing and Population Genomics in non-model organisms. *Trends in Ecology & Evolution* 29, no. 1 (2014)
- Endler JA. Natural Selection in the Wild. *Princeton University Press.* (1986)
- Eyre-Walker A. Changing effective population size and the McDonald-Kreitman test. *Genetics*, 162(4):2017-2024 (2002)
- Eyre-Walker A and Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution* 26: 2097-2108. (2009)
- Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J mol evol.* 17(6):368-76 (1981)
- Fay JC and Wu CI. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3): 1405-1413 (2000)
- Fay JC, Wyckoff GJ and Wu CI. Positive and negative selection on the human genome. *Genetics* 158(3), 1227-1234 (2001)
- Feuk L, Carson AR and Scherer SW. Structural variation in the human genome. *Nature Review Genetics* 7(2): 85-97 (2006)
- Fiston-Lavier AS, Singh ND, Lipatov M and Petrov DA. *Drosophila melanogaster* recombination rate calculator. *Gene* 463, 18-20 (2010)
- Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32: W273-W279 (2004)
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM et al (The international HapMap consortium). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851-861 (2007)
- Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW and Lee C. Copy number variation: New insights in genome diversity. *Genome Research* 16: 949-961 (2006)
- Fu YX and Li WH. Statistical tests of neutrality of mutations. *Genetics*, 133(3): 693-709 (1993)
- Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147, 915-925 (1997)
- García-Dorado A. Understanding and Predicting the Fitness Decline of Shrunk Populations: Inbreeding, Purging, Mutation, and Standard Selection. *Genetics* 190(4): 1461-1476 (2012)
- Gillespie JH. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 155(2): 909-919 (2000a)
- Gillespie JH. The neutral theory in an infinite population. *Genetics* 261(1): 11-18. (2000b)
- Gillespie JH. Is the population size of a species relevant to its evolution? *Evolution Int J Org Evolution* 55(11): 2161-2169 (2001)
- Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Bio.* 25, R86 (2010)
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B et al. Life with 6000 genes. *Science* 274(5287): 546, 563-547 (1996)
- Gossman TI et al. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution* 27(8) 1822-1832 (2012)
- Greenblatt MS, Grollman AP and Harris CC. Deletions and insertions in the p53 tumor suppressor gene in human cancers: Confirmation of the DNA polymerase slippage/misalignment model. *Cancer Res* 56: 2130-2136 (1996)
- Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *PathoGenetics* 1:4 (2008)
- Hahn MW. Towards a selection theory of molecular evolution. *Evolution* 62-2: 255-265 (2008)

- Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics* 43: 269–276 (2011)
- Hardy GH. Mendelian Proportions in a Mixed Population. *Science* 28 (706): 49–50 (1908)
- Harris H. Enzyme polymorphisms in man. *Proc R Soc Lond B Biol Sci* 164(995): 298-310 (1966)
- Hartl DL and Clark AG. *Principles of Population Genetics*. Sinauer Associates Inc. Sunderland, Massachusetts. (1997)
- Hastings PJ, Ira G and Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet* 5: e1000327 (2009)
- Hershberg R and Petrov D. Global rules for optimal codon choice. *PLoS Genetics*, 5, e1000556. doi: 10.1371/journal.pgen.1000556 (2009)
- Hill WG and Robertson A. The effect of linkage on limits to artificial selection. *Genetical Research* 8: 269–294 (1966)
- Hill WG and Robertson A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38: 226-231 (1968)
- Hodgkinson A and Eyre-Walker A. Variation in the mutation rate across mammalian genomes. *Nature Review Genetics* 12: 756-766 (2011)
- Holt C and Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491 (2011)
- Hu TT, Eisen MB, Thornton KR and Andolfatto P. A second generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Research* 23, 89–98. (2013)
- Huang W, Massouras A *et al.* Natural Variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Research* 24, 1193-1208 (2014)
- Hubbard, T, Barker, D, Birney, E, Cameron, G, Chen, Y, Clark, L, Cox, T, Cuff, J, Curwen, V, Down, T *et al.* The Ensembl genome database project. *Nucleic acids research* 30, 38-41 (2002)
- Hudson RR and Kaplan NL. Deleterious background selection with recombination. *Genetics* 141: 1605–1617 (1995)
- Hudson RR, Kreitman M and Aguade M. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116(1): 153-159 (1987)
- Hutter S, Vilella AJ and Rozas J. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics* 7, 409 (2006)
- Johnson FM, Kanapi CG, Richardson RH, Wheeler MR and Stone WS. An analysis of polymorphisms among isozyme loci in dark and light *Drosophila ananassae* strains from American and Western Samoa. *Proc Natl Acad Sci USA*, 56(1): 119-125(1966)
- Jovelin R and Cutter AD. Fine-scale signatures of molecular evolution reconcile models of indel-associated mutation. *Genome Biology and Evolution* 5:978-986 (2013)
- Jukes TH and Cantor CR. Mammalian protein metabolism. *Evolution of Protein Molecules*, pp. 21-132, edited by H. N. Munro. Academic Press, New York (1969)
- Kaplan LN, Hudson RR, Langley CH. The "Hitchhiking effect" Revisited. *Genetics* 123:887-899 (1989)
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30, 772-780 (2013)
- Keightley PD and Eyre-Walker A. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Phil. Trans. R. Soc. B* 365 (2010)
- Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S and Blaxter ML. Analysis of the Genome Sequences of Three *Drosophila Melanogaster* Spontaneous Mutation Accumulation Lines Analysis of the Genome Sequences of Three *Drosophila Melanogaster* Spontaneous Mutation Accumulation Lines. *Genome Research* 19(7), 1195–1201 (2009)
- Keller A. *Drosophila melanogaster's* history as a human commensal. *Current Biology* 17: R77–R81 (2007)

- Kelly JK. A test of neutrality based on interlocus associations. *Genetics*, 146(3): 1197-1206 (1997)
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D: The human genome browser at UCSC. *Genome research* 2002, 12, 996-1006
- Kida Y, Maeda Y, Shiraishi T, Suzuki T and Ogura T. Chick Dach1 interacts with the Smad complex and Sin3a to control AER formation and limb development along the proximodistal axis. *Development*, 131(17): 4179-4187 (2004)
- Kidwell MG. Lateral transfer in natural populations of eukaryotes. *Annu Rev Genet* 27, 236–256 (1993)
- Kimura M. Evolutionary rate at the molecular level. *Nature*, 217: 624-626 (1968)
- Kimura M. The neutral theory of molecular evolution. *Cambridge* (1983)
- King JL and Jukes TH. Non-Darwinian evolution. *Science*, 164(3881): 788-98 (1969)
- Kingman JFC. Origins of the Coalescent: 1974-1982. *Genetics* 156(4), 1461-1463 (2000)
- Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C and Schlötterer C. PoPoolation, a Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS One* 6(1): e15925. doi:10.1371/journal.pone.0015925 (2011)
- Kulathinal RJ, Bennett SM, Fitzpatrick CL and Noor MAF. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *PNAS* 105(29), 10051-10056 (2008)
- Kumar S, Tamura K, Jakobsen IB and Nei M. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* 17, 1244-1245 (2001)
- Lachaise D, Cariou ML, David JR, Lemeunier F, Tsacas L and Asburner M. Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evolutionary Biology* 22: 159–225 (1988)
- Lamesch P, Berardini TZ, Li D, Swarbreck D *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* doi: 10.1093/nar/gkr1090 (2011)
- Langley CH, Macdonald J, Miyashita N and Aguade M. Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans*. *Proc Natl Acad Sci USA* 90(5): 1800-1803 (1993)
- Langley CH, Crepeau M, Cardeno C, Corbett-Detig R, and Stevens K. Circumventing Heterozygosity: Sequencing the Amplified Genome of a Single Haploid *Drosophila melanogaster* Embryo. *Genetics* 188, 239–246 (2011)
- Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, *et al.* Genomic Variation in Natural Populations of *Drosophila Melanogaster*. *Genetics* 192(2), 533–98 (2012)
- Lawrie DS, Messer PW, Hershberg R, Petrov D. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genetics* DOI: 10.1371/journal.pgen.1003527 (2013)
- Lee E, Harris N, Gibson M, Chetty R and Lewis S. Apollo: a community resource for genome annotation editing. *Bioinformatics* 25, 1836-7 (2009)
- Lee JA, Carvalho CM, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, 131: 1235–1247 (2007)
- Leushkin EV, Bazykin GA, Kondrashov AS. Strong mutational bias toward deletions in the *Drosophila melanogaster* genome is compensated by selection. *Genome Biology and evolution* 5, 514-524 (2013)
- Levinson G and Gutman GA. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol Biol Evol*, 4: 203–221 (1987)
- Lewontin RC and Hubby JL. 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 (1966)
- Lewontin RC and Kojima K. The evolutionary dynamics of complex polymorphisms. *Evolution* 14:458-472 (1960)
- Lewontin RC and Krakauer J. Distribution of gene frequency as a test of the theory of the selective



- neutrality of polymorphisms. *Genetics* 74, 175-195 (1973)
- Lewontin RC. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49: 49-67 (1964)
- Lewontin RC. The units of selection. *Annual Review of Ecology and Systematics*. 1: 1-18 (1970)
- Lewontin RC. The genetic basis of evolutionary change. Columbia University Press, New York. (1974)
- Lewontin RC. Biology as Ideology: The Doctrine of DNA. Anansi Press / Stoddard Publishing Company (1992)
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25, 2078-9 (2009)
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589–595 (2010)
- Li WH, Wu CI and Luo CC. 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(2): 150-174 (1985)
- Librado P and Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451–1452 (2009)
- Lintner JA. *First Annual Report on the Injurious and Other Insects of the State of New York*. Weed, Parsons, Albany, NY (1882)
- Liti G, Carter DM, Moses AM *et al*. Population genomics of domestic and wild yeasts. *Nature* 458, 337-341 (2009)
- Mackay *et al*. The *Drosophila melanogaster* Genetic Reference Panel. *Nature* 482, 173–178 (2012)
- Mackay T, Richards S, Gibbs R. Proposal to Sequence a *Drosophila* Genetic Reference Panel: A Community Resource for the Study of Genotypic and Phenotypic Variation. White paper (2008)
- Marais G, Mouchiroud D and Duret L. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc Natl Acad Sci USA* 98, 5688-5692 (2001)
- Martin-Campos JM, Comeron JM, Miyashita N and Aguade M. Intraspecific and interspecific variation at the y-ac-sc region of *Drosophila simulans* and *Drosophila melanogaster*. *Genetics*, 130(4): 805-816 (1992)
- Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond JR, Strelets VB, Wilson RJ, the FlyBase consortium. FlyBase: improvements to the bibliography. *Nucleic Acids Research* 41, D751–D757 (2013)
- Massouras A, Hens K, Gubelmann C, Uplekar S, Decouttere F, Rougemont J, Cole ST, Deplancke B. Primer-initiated sequence synthesis to detect and assemble structural variants. *Nature Methods* 7: 485–486 (2010)
- Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W, Ayroles JF, Dermitzakis ET, Stone EA, Jensen JD, Mackay TFC, *et al*. Genomic variation and its impact on gene expression in *Drosophila melanogaster*. *PLoS Genet* 8: e1003055 (2012)
- Maxam AM, Gilbert W. A new method for sequencing DNA. *PNAS* 74 (2): 560–4 (1977)
- Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS and Dubchak I. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16, 1046–1047 (2000)
- Mayr E. *Systematics and the Origin of Species*. Columbia University Press, New York (1942)
- McGinn S, Gut IG. DNA sequencing – spanning the generations. *New biotechnology* 30(4), 366-372 (2013)
- McDonald JH and Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351(6328): 652-654 (1991)
- McDonald MJ, Wang WC, Huang HD and Leu JY. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol* 9: e1000622 (2011)
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler

- D, Gabriel S, Daly M, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297–1303 (2010)
- McVey M, Larocque JR, Adams MD and Sekelsky JJ. Formation of deletions during double-strand break repair in *Drosophila* DmBlm mutants occurs after strand invasion. *Proc Natl Acad Sci* 101: 15694–15699 (2004)
- Medina I, Salavert F, Sanchez R, De Maria A, Alonso R, Escobar R, Bleda M, Dopazo J. Genome Maps, a new generation genome browser. *Nucleic Acids Research* 41, W41-W46 (2013)
- Meisel RP and Connallon T. The faster-X effect: integrating theory and data. *Trends in Genetics*, 29(9): 537-544 (2013)
- Messer PW and Petrov DA. Frequent Adaptation and the McDonald-Kreitman Test. *PNAS* 110(21), 8615–20 (2013)
- Miyashita N and Langley CH. Molecular and phenotypic variation of the white locus region in *Drosophila melanogaster*. *Genetics* 120(1): 199-212 (1988)
- Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, *et al.* The origin, evolution and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* 23: 749–761. (2013)
- Morgan TH, Sturtevant AH, Muller HJ and Bridges CB. The Mechanism of Mendelian Heredity. Henry Holt and Company, New York. (1915)
- Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet.* 19:R131-R6 (2010)
- Mungall CJ, Emmert DB and the FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* 23, 337-346 (2007)
- Neher RA., Kessinger TA and Shraiman BI. Coalescence and Genetic Diversity in Sexual Populations under Selection. *PNAS*, 110(39): 15836–41 (2013)
- Nei M. *Molecular evolutionary genetics*. Columbia University Press, New York. (1987)
- Nei M and Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3(5): 418-426 (1986)
- Nevo E, Beiles A and Ben-Shlomo R. The evolutionary significance of genetic diversity: ecological, demographic and life history correlates. *Lecture notes in biomathematics*, pp. 13-213, S. Levin, Ed., vol. 53, Evolutionary dynamics of genetic diversity. G. S. Mani, Ed. (Springer-Verlag), Berlin (1984)
- Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of Next-Generation Sequencing Technologies. *Analytical chemistry* 83, 4327-4341 (2011)
- Nielsen R. Molecular signatures of Natural Selection. *Annu Rev Genet* 39, 197-218 (2005)
- Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG and Bustamante C. Genomic scans for selective sweeps using SNP data. *Genome Research* 15, 1566-1575 (2005)
- Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936 (1998)
- Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for multiple sequence alignments. *JMB* 302, 205-217 (2000)
- Ohta T. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J Mol Evol* 40(1): 56-63 (1995)
- Oleksyk TK, Zhao K, De La Vega FM, Gilbert DA, O'Brien SJ and Smith MW. Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS One* 3, e1712 (2008)
- Olekysk TK, Smith MW and O'Brien SJ. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of The Royal Society B.* 365: 185-205 (2009)
- Ometto L, Stephan W and De Lorenzo D. Insertion/deletion and nucleotide

- polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* 169, 1521–1527 (2005)
- Ovcharenko I, Nobrega MA, Loots GG and Stubbs L. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res.* 32, W280–W286 (2004)
- Pang AW, Migita O, Macdonald JR, Feuk L, Scherer SW. Mechanisms of formation of structural variation in a fully sequenced human genome. *Human mutation* 34(2), 345–354 (2013)
- Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM and Andolfatto P. On the utility of short intron sequences as a reference for detection of positive and negative selection in *Drosophila*. *Molecular Biology and Evolution* 27(6):1226–1234 (2010)
- Petrov DA. Mutational equilibrium model of genome size evolution. *Theor Popul Biol* 61, 533–546 (2002)
- Piganeau G & Eyre-Walker A. Estimating the distribution of fitness effects from DNA sequence data: Implications for the molecular clock. *PNAS* 100(18), 10335–10340 (2003)
- Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchon P, *et al.* Population Genomics of Sub-Saharan *Drosophila Melanogaster*: African Diversity and Non-African Admixture. *PLoS Genetics* 8(12), e1003080 (2012)
- Pool JE, Hellmann I, Jensen JD and Nielsen R. Population genetic inference from genomic sequence variation. *Genome research* 20: 291–300 (2010)
- Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28: i333–i339 (2012)
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH *et al.* Global variation in copy number in the human genome. *Nature* 444(7118), 444–454 (2006)
- Roberts DB. *Drosophila melanogaster*: the model organism. *Entomologia Experimentalis et Applicata* 121:93–103 (2006)
- Romanes GJ. Post-Darwinian Questions: Heredity and Utility. *The open court publishing company* (1906, first published 1895)
- Sabeti PC, Reich DE, Higgins JM *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837 (2002)
- Sabeti PC *et al.* The case for selection at CCR5-Delta32. *PLoS Biol* 3, e378 (2005)
- Sabeti PC, Varilly P, Fry B, Lohmueller J *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918 (2007)
- Sackton TB *et al.* Population genomic inferences from sparse high-throughput sequencing of two populations of *Drosophila melanogaster*. *Genome Biol. Evol* 1, 449–465 (2009)
- Sanger F and Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94 (3): 441–8 (1975)
- Schattner P. *Genomes, Browsers & Databases. Data-Mining Tools for Integrated Genomic Databases.* Cambridge University Press, New York (2008)
- Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive Natural Selection in the *Drosophila* genome? *PLoS Genetics* 5(6): e1000495. doi: 10.1371/journal.pgen.1000495 (2009)
- Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* 20: 273–280 (2010)
- Schrider DR, Houle D, Lynch M and Hahn MW. Rates and Genomic Consequences of Spontaneous Mutational Events in *Drosophila Melanogaster*. *Genetics* 194(4), 937–954 (2013)
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* 15, 1034–1050 (2005)

- Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD and Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. 7:539 (2011)
- Simpson GG. *Tempo and Mode in Evolution*. Columbia University Press, New York (1944)
- Smith CD, Shu S, Mungall CJ and Karpen GH. The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. *Science* 316, 1586-1591 (2007)
- Smith DR, Quinlan AR., Peckham HE *et al.* Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research* 18(10), 1638–1642 (2008)
- Smith JM and Haigh J. The hitch-hiking effect of a favourable gene. *Genet Res* 23(1): 23-35. (1974)
- Smith MW and O'Brien SJ. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat. Rev. Genet.* 6, 623–632 (2005).
- Smith NGC and Eyre-Walker A. Adaptive protein evolution in “*Drosophila*”. *Nature* 415, 1022-1024 (2002)
- St. Pierre SE, Ponting L, Stefancsik R, McQuilton P, and the FlyBase Consortium. FlyBase 102 - advanced approaches to interrogating FlyBase. *Nucleic Acids Res* doi: 10.1093/nar/gkt1092 (2014)
- Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehtväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, and Birney E. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research* 12, 1611-8 (2002)
- Stebbins GL. *Variation and Evolution in Plants*. Columbia University Press, New York (1950)
- Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al.* The generic genome browser: a building block for a model organism system database. *Genome research* 12, 1599-1610 (2002)
- Stein LD. Genome annotation: from sequence to biology. *Nature review genetics* 2, 493-503 (2001)
- Stephan W and Langley CH. Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the vermilion and forked loci. *Genetics* 121(1): 89-99 (1989)
- Stephan W and Li H. The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* 98(2):65-8. (2007)
- Stephan W and Mitchell SJ. Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. *Genetics*, 132(4): 1039-1045 (1992)
- Stevenson LS, and Noor MAF. Genetic and evolutionary correlates of fine-scale recombination rate variation in *Drosophila persimilis*. *Journal of Molecular Evolution*, 71: 332-345 (2010)
- Stoletzki N and Eyre-Walker A. Estimation of the Neutrality Index. *Molecular Biology and Evolution* 28(1), 63-70 (2011)
- Stone EA. Joint genotyping on the fly: Identifying variation among a sequenced panel of inbred lines. *Genome research* 22: 966-974 (2012)
- Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E and Inouye M. Frameshift mutations and the genetic code. *Cold Spring Harb Symp Quant Biol*, 31: 77–84 (1966)
- Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105(2): 437-460 (1983)
- Tajima F. Measurement of DNA polymorphism in *Mechanisms of molecular evolution*, edited by N. Takahata and A. G. Clark. Sinauer Associates Inc., Sunderland, Massachusetts. (1993)
- Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3): 585-595 (1989)
- Tajima F. The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics* 143(3):1457-1465 (1996)
- Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintrón W, Burchard E, Gonz I and Risch NJ. Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am. J. Hum. Genet.* 81, 626–633 (2007)

- Taylor MS, Ponting CP and Copley RR. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Res* 14: 555–566 (2004)
- The International HapMap Consortium: The International HapMap Project. *Nature* 426, 789-796 (2003)
- Thornton K, Bachtrog D and Andolfatto P. X chromosomes and autosomes evolve at similar rates in *Drosophila*: No evidence for faster-X protein evolution. *Genome Research* 16:498-504 (2006)
- Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J and Chen JQ. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 455: 105-108 (2008)
- Väli U, Brandstrom M, Johansson M, Ellegren H. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genet.* 9:8 (2008)
- Vicoso B and Charlesworth B. Evolution on the X chromosome: unusual patterns and processes. *Nature Review Genetics* 7:645-653 (2006)
- Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, Mesirov JP, Satoh N, Satou Y, Nusbaum C, Birren B, Galagan J and Lander E. Assembly of polymorphic genomes: Algorithms and application to *Ciona savignyi*. *Genome Research* 15: 1127-1135 (2005)
- Voight BF, Kudravalli S, Wen X, Pritchard J. A map of recent positive selection in the human genome. *PLoS Biology* 4, e154 (2006)
- Watson JD and Crick FHC. A structure for deoxyribose nucleic acid. *Nature*, 171: 737-738 (1953)
- Watterson GA. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7(2): 256-276 (1975)
- Weinberg W. Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg* 64: 368–382 (1908)
- Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872-876 (2008)
- Wright S. Evolution in Mendelian populations. *Genetics* 16: 97-159 (1931)
- Yang W and Woodgate R. What a difference a decade makes: Insights into translesion DNA synthesis. *Proc Natl Acad Sci USA* 104: 15591–15598 (2007)
- Yang Z and Bielawski JP. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, 15(12): 496-503 (2000)
- Yang Z and Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol* 46, 409–418 (1998)
- Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25: 2865–2871 (2009)
- Yook K, Harris TW, Bieri T, Cabunoc A, Chan J and the WormBase consortium. WormBase 2012: more genomes, more data, new website. *Nucleic Acid Research* 40, D735-D741 (2012)
- Zhao H, Yang Y, Lin H, Zhang X, Mort M, Cooper DN, Liu Y, Zhou Y. DDIG-in: discriminating between disease-associated non-frameshifting micro-indels. *Genome biology* 14:R23 (2013)
- Zarrei M, MacDonald JR, Merico D & Scherer SW. A copy number variation map of the human genome. *Nature reviews. Genetics* 16, 172-83 (2015)
- Zuckerkindl E and Pauling L. Molecular disease, evolution, and genetic heterogeneity. *Horizons in Biochemistry*, pp. 189-225, edited by M. Kasha and B. Pullman. Academic Press, New York (1962)



---

## **Annex A**

### **PopDrowser: the Population Drosophila Browser**

## PopDrowser: the Population *Drosophila* Browser

Miquel Ràmia<sup>1,†</sup>, Pablo Librado<sup>2,†</sup>, Sònia Casillas<sup>1,†</sup>, Julio Rozas<sup>2</sup> and Antonio Barbadilla<sup>1,\*</sup>

<sup>1</sup>Institut de Biotecnologia i de Biomedicina and Departament de Genètica i de Microbiologia (Facultat de Biociències), Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona) and <sup>2</sup>Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

Associate Editor: Jeffrey Barret

### ABSTRACT

**Motivation:** The completion of 168 genome sequences from a single population of *Drosophila melanogaster* provides a global view of genomic variation and an understanding of the evolutionary forces shaping the patterns of DNA polymorphism and divergence along the genome.

**Results:** We present the 'Population *Drosophila* Browser' (PopDrowser), a new genome browser specially designed for the automatic analysis and representation of genetic variation across the *D. melanogaster* genome sequence. PopDrowser allows estimating and visualizing the values of a number of DNA polymorphism and divergence summary statistics, linkage disequilibrium parameters and several neutrality tests. PopDrowser also allows performing custom analyses on-the-fly using user-selected parameters.

**Availability:** PopDrowser is freely available from <http://PopDrowser.uab.cat>.

**Contact:** [miquel.ramia@uab.cat](mailto:miquel.ramia@uab.cat)

Received on October 6, 2011; revised on November 29, 2011; accepted on December 8, 2011

## 1 INTRODUCTION

Population genetics studies have been so far based on fragmentary and non-random samples of genomes, providing a partial and often biased view of the population genetics processes (Begun *et al.*, 2007). A new dimension to genetic variation studies is provided by the new availability of within-species genomes. Next-generation sequencing technologies are making affordable genome-wide population genetics data, not only for humans and the main model organisms, but also for most organisms on which research is actively carried out on genetics, ecology or evolution (Pool *et al.*, 2010).

Genome browsers are very useful tools to query and visualize disparate annotations at different genomic locations using a web user interface (Schattner, 2008). A number of web-based genome browsers displaying genetic variation data are already available (Benson *et al.*, 2002; Dubchak and Ryaboy, 2006; Frazer *et al.*, 2007; Hubbard *et al.*, 2002; Kent *et al.*, 2002; Stein *et al.*, 2002). Such browsers, however, are not well suited to deal with population genomics sequence information. For example, HapMap (International HapMap Consortium, 2003), the most comprehensive

genome browser of variation data so far, contains information on single nucleotide polymorphisms (SNPs), Copy Number Variations (CNVs) and linkage disequilibrium of human populations. It does not offer, however, genetic variation estimates along sliding-windows or neutrality-based tests.

The *Drosophila* Genetic Reference Panel (DGRP) (TMackay *et al.*, accepted for publication) has recently sequenced and analyzed the patterns of genome variation in 168 inbred lines of *Drosophila melanogaster* from a single population of Raleigh (USA), and conducted a genome-wide association analysis of some phenotypic traits. A major goal of this project is to create a resource of common genetic polymorphism data to aid further population genomics analyses. As a part of this DGRP project, here we present a modified Gbrowse specifically designed for the automatic estimation and representation of population genetic variation in *D. melanogaster*, the 'Population *Drosophila* Browser' (PopDrowser). Unlike other population analysis tools (Hutter *et al.*, 2006; Kofler *et al.*, 2011), the PopDrowser is a genome browser, which can be customized to create analogous resources for any other species with within-species polymorphism data.

## 2 IMPLEMENTATION

### 2.1 Input data

The initial input data are a set of 168 aligned intraspecific *D. melanogaster* sequences from the DGRP project, and also include the genome sequences of *Drosophila yakuba* and *Drosophila simulans*, which were used as outgroup species.

### 2.2 Interface and implementation

PopDrowser allows reporting precomputed estimates of several DNA variation measures along each chromosome arm through the combined implementation of the programs PDA 2 (Casillas and Barbadilla, 2006), MKT (Egea *et al.*, 2008) and VariScan 2 (Hutter *et al.*, 2006). The data and summary statistics are graphically displayed along the chromosome arms on a web-based user interface using the Gbrowse software.

PopDrowser also includes an innovative capability that allows performing custom analyses on-the-fly. After selecting a chromosome region and a particular track, the user can conduct exhaustive analyses by defining their own custom input parameters. Furthermore, users can choose to either visualize the output of their analyses graphically in the browser—as a new track—or to

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.



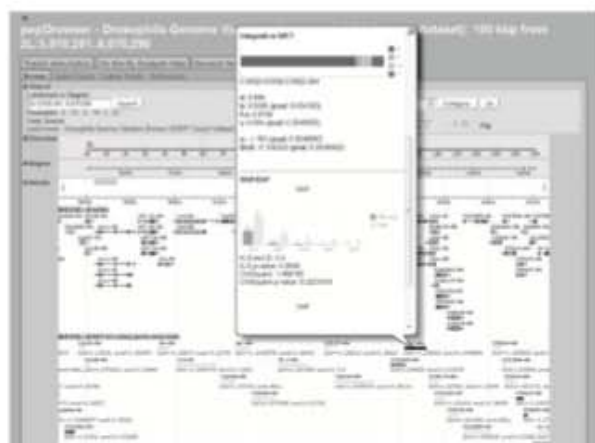


Fig. 1. PopDrowser snapshot showing the results of the McDonald–Kreitman tests in the *ade2-RA* gene within its genome context.

download it in a tabulated text file. The estimates available for on-the-fly analyses are specified in Table 1.

The current implementation is running in an Ubuntu 10.04 Linux x64 server, 2 IntelXeon 3Ghz processors, 32GB RAM with Apache.

### 2.3 Output

Along with reference genome annotations, the genome browser output includes measures of a number of nucleotide summary statistics, such as the levels of nucleotide diversity ( $\pi$  and  $\theta$ ), DNA divergence between species ( $K$ ), different measures of linkage disequilibrium and genome-wide neutrality tests. Such analyses are computed along each chromosome arm in non-overlapping sliding windows of 0.05, 0.1, 0.5, 1, 10, 50 and 100 kb. For each gene, the browser also provides a single track including information of the generalized and the integrative McDonald–Kreitman tests (McDonald and Kreitman, 1991; T.Mackay *et al.*, accepted for publication) along with minor and derived allele frequency (MAF, DAF) spectrums (Fig. 1). All the tracks included in the PopDrowser are summarized in Table 1.

### ACKNOWLEDGEMENTS

We thank Raquel Egea for helping in implementing the MKT software and Albert Vernon Smith from HapMap for his help in the implementation of the pie graph glyph. We also thank Dave Clements, Lincoln Stein and Scott Cain for technical support, and John H. Werren for valuable discussions on the browser. This paper was prepared with full knowledge and support of the DGRP Consortium.

**Funding:** Ministerio de Ciencia e Innovación (Spain) (BFU2009-09504 to A.B., BFU2010-15484 to J.R.); Catalanian Comissió Interdepartamental de Recerca i Innovació Tecnològica (2009SGR-88 to A. Ruiz; 2009SGR-1287 to M. Aguadé); Departament de Genètica i de Microbiologia of the Universitat Autònoma de Barcelona (409-04-2/08 to M.R.); ICREA Academia (Generalitat de Catalunya to J.R., in part).

**Conflict of Interest:** none declared.

Table 1. Summary of the PopDrowser tracks

Category	Gene annotations and estimates
<i>Drosophila melanogaster</i> reference annotations (build 5.13) and recombination	Gene structure, mRNA, CDS (Coding Sequence), ncRNA, tRNA, orthologous genes, phastCons, GC content, local recombination rate (Fiston-Lavier <i>et al.</i> , 2010)
Density tracks	Genes, microsatellites, transposons, CDS, SNPs
Nucleotide variants	SNPs, single nucleotide fixations
Measures of nucleotide variation and LD <sup>a</sup>	Number of segregating sites ( $S$ ), total minimum number of mutations ( $\eta$ ), number of singletons ( $\eta_e$ ), nucleotide diversity ( $\pi$ ), Watterson's estimator of nucleotide diversity per site ( $\theta$ ), number of haplotypes ( $h$ ), haplotype diversity ( $Hd$ ), nucleotide divergence per site (corrected by Jukes–Cantor) ( $K$ ). LD: $D$ , absolute $D$ ( $ D $ ), $D'$ , absolute $D'$ ( $ D' $ ), $r^2$
Neutrality tests <sup>a</sup>	Fu and Li's $D$ , $D^a$ , $F$ , $F^a$ , Fay and Wu's $H$ , Tajima's $D$ , Fu's $F_S$ statistics. MKT (per gene)

LD, linkage disequilibrium; CDS, coding sequence.

<sup>a</sup>Estimates available for on-the-fly analyses (except MKT per gene).

### REFERENCES

- Begun, D.J. *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.*, **5**, e310.
- Benson, D.A. *et al.* (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
- Casillas, S. and Barbado, A. (2006) PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. *Nucleic Acids Res.*, **34**, W632–W634.
- Dubchak, I. and Ryaboy, D.V. (2006) VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. *Methods Mol. Biol.*, **2006**, 338, 69–89.
- Egea, R. *et al.* (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.
- Fiston-Lavier, A.S. *et al.* (2010) *Drosophila melanogaster* recombination rate calculator. *Gene*, **463**, 18–20.
- Frazer, K.A. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
- Hubbard, T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
- Hutter, S. *et al.* (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.
- International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.
- Kent, W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
- Kofler, R. *et al.* (2011) PoPoolation, a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One*, **6**, e15925.
- McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, **351**, 652–654.
- Pool, J.E. *et al.* (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.
- Schattner, P. (2008) *Genomes, Browsers and Databases. Data-Mining Tools for Integrated Genomic Databases*. Cambridge University Press, New York.
- Stein, L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.



---

## **Annex B**

The *Drosophila melanogaster* genetic reference panel

# The *Drosophila melanogaster* Genetic Reference Panel

Trudy F. C. Mackay<sup>1\*</sup>, Stephen Richards<sup>2\*</sup>, Eric A. Stone<sup>1\*</sup>, Antonio Barbadilla<sup>3\*</sup>, Julien F. Ayroles<sup>1†</sup>, Dianhui Zhu<sup>2</sup>, Sònia Casillas<sup>3†</sup>, Yi Han<sup>2</sup>, Michael M. Magwire<sup>1</sup>, Julie M. Cridland<sup>4</sup>, Mark F. Richardson<sup>5</sup>, Robert R. H. Anholt<sup>6</sup>, Maite Barrón<sup>3</sup>, Crystal Bess<sup>2</sup>, Kerstin Petra Blankenburg<sup>2</sup>, Mary Anna Carbone<sup>1</sup>, David Castellano<sup>3</sup>, Lesley Chaboub<sup>2</sup>, Laura Duncan<sup>1</sup>, Zeke Harris<sup>1</sup>, Mehwish Javaid<sup>2</sup>, Joy Christina Jayaseelan<sup>2</sup>, Shalini N. Jhangiani<sup>2</sup>, Katherine W. Jordan<sup>1</sup>, Fremiet Lara<sup>2</sup>, Faye Lawrence<sup>1</sup>, Sandra L. Lee<sup>2</sup>, Pablo Librado<sup>7</sup>, Raquel S. Linheiro<sup>5</sup>, Richard F. Lyman<sup>1</sup>, Aaron J. Mackey<sup>8</sup>, Mala Munidasa<sup>2</sup>, Donna Marie Muzny<sup>2</sup>, Lynne Nazareth<sup>2</sup>, Irene Newsham<sup>3</sup>, Lora Perales<sup>2</sup>, Ling-Ling Pu<sup>2</sup>, Carson Qu<sup>2</sup>, Miquel Ràmia<sup>3</sup>, Jeffrey G. Reid<sup>2</sup>, Stephanie M. Rollmann<sup>1†</sup>, Julio Rozas<sup>7</sup>, Nehad Saada<sup>2</sup>, Lavanya Turlapati<sup>1</sup>, Kim C. Worley<sup>2</sup>, Yuan-Qing Wu<sup>2</sup>, Akihiko Yamamoto<sup>1</sup>, Yiming Zhu<sup>2</sup>, Casey M. Bergman<sup>5</sup>, Kevin R. Thornton<sup>4</sup>, David Mittelman<sup>9</sup> & Richard A. Gibbs<sup>2</sup>

A major challenge of biology is understanding the relationship between molecular genetic variation and variation in quantitative traits, including fitness. This relationship determines our ability to predict phenotypes from genotypes and to understand how evolutionary forces shape variation within and between species. Previous efforts to dissect the genotype-phenotype map were based on incomplete genotypic information. Here, we describe the *Drosophila melanogaster* Genetic Reference Panel (DGRP), a community resource for analysis of population genomics and quantitative traits. The DGRP consists of fully sequenced inbred lines derived from a natural population. Population genomic analyses reveal reduced polymorphism in centromeric autosomal regions and the X chromosome, evidence for positive and negative selection, and rapid evolution of the X chromosome. Many variants in novel genes, most at low frequency, are associated with quantitative traits and explain a large fraction of the phenotypic variance. The DGRP facilitates genotype-phenotype mapping using the power of *Drosophila* genetics.

Understanding how molecular variation maps to phenotypic variation for quantitative traits is central for understanding evolution, animal and plant breeding, and personalized medicine<sup>1,2</sup>. The principles of mapping quantitative trait loci (QTLs) by linkage to, or association with, marker loci are conceptually simple<sup>1,2</sup>. However, we have not yet achieved our goal of explaining genetic variation for quantitative traits in terms of the underlying genes; additive, epistatic and pleiotropic effects as well as phenotypic plasticity of segregating alleles; and the molecular nature, population frequency and evolutionary dynamics of causal variants. Efforts to dissect the genotype-phenotype map in model organisms<sup>3,4</sup> and humans<sup>5-7</sup> have revealed unexpected complexities, implicating many, novel loci, pervasive pleiotropy, and context-dependent effects.

Model organism reference populations of inbred strains that can be shared among laboratories studying diverse phenotypes, and for which environmental conditions can be controlled and manipulated, greatly facilitate efforts to dissect the genetic architecture of quantitative traits<sup>3,4</sup>. Measuring many individuals of the same homozygous genotype increases the accuracy of the estimates of genotypic value<sup>1</sup> and the power to detect variants, and genotypes of molecular markers need only be obtained once. We constructed the *Drosophila melanogaster* Genetic Reference Panel (DGRP) as such a community resource. Unlike previous populations of recombinant inbred lines derived from limited samples of genetic variation, the DGRP consists

of 192 inbred strains derived from a single outbred population. The DGRP contains a representative sample of naturally segregating genetic variation, has an ultra-fine-grained recombination map suitable for precise localization of causal variants, and has almost complete euchromatic sequence information.

Here, we describe molecular and phenotypic variation in 168 resequenced lines comprising Freeze 1.0 of the DGRP, population genomic inferences of patterns of polymorphism and divergence and their correlation with genomic features, local recombination rate and selection acting on this population, genome-wide association mapping analyses for three quantitative traits, and tools facilitating the use of this resource.

## Molecular variation in the DGRP

We constructed the DGRP by collecting mated females from the Raleigh, North Carolina, USA, population, followed by 20 generations of full-sibling inbreeding of their progeny. We sequenced 168 DGRP lines using a combination of Illumina and 454 sequencing technology: 29 of the lines were sequenced using both platforms, 129 lines have only Illumina sequence, and 10 lines have only 454 sequence. We mapped sequence reads to the *D. melanogaster* reference genome, re-calibrated base quality scores, and locally re-aligned Illumina reads. Mean sequence coverage was 21.4× per line for Illumina sequences and 12.1× per line for 454 sequences (Supplementary

<sup>1</sup>Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695, USA. <sup>2</sup>Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030 USA. <sup>3</sup>Genomics, Bioinformatics and Evolution Group, Institut de Biociologia i de Biomedicina - IBB/Department of Genetics and Microbiology, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain. <sup>4</sup>Department of Ecology and Evolutionary Biology, University of California - Irvine, Irvine, California 92697, USA. <sup>5</sup>Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. <sup>6</sup>Department of Biology, North Carolina State University, Raleigh, North Carolina 27695, USA. <sup>7</sup>Molecular Evolutionary Genetics Group, Department of Genetics, Faculty of Biology, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain. <sup>8</sup>Center for Public Health Genomics, University of Virginia, PO Box 800717, Charlottesville, Virginia 22908, USA. <sup>9</sup>Virginia Bioinformatics Institute and Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia 24061, USA. <sup>†</sup>Present addresses: FAS Society of Fellows, Harvard University, 78 Mt Auburn Street, Cambridge, Massachusetts 02138, USA (J.F.A.); Functional Comparative Genomics Group, Institut de Biociologia i de Biomedicina - IBB, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain (S.C.); Department of Biological Sciences, University of Cincinnati, Cincinnati, Ohio 45221, USA (S.M.R.).

\*These authors contributed equally to this work.

Table 1). On average, we assayed 113.5 megabases (94.25%) of the euchromatic reference sequence with ~22,000 read mapping gaps per line (Supplementary Table 2). We called 4,672,297 single nucleotide polymorphisms (SNPs) using the Joint Genotyper for Inbred Lines (JGIL; E.A.S., personal communication), which takes into account coverage and quality sequencing statistics, and expected allele frequencies after 20 generations of inbreeding from an outbred population initially in Hardy–Weinberg equilibrium. In cases where base calls were made by both technologies, concordance was 99.36% (Supplementary Table 3).

The SNP site frequency distribution (Fig. 1a) is characterized by a majority of low frequency variants. The numbers of SNPs vary by chromosome and site class (Fig. 1b). Linkage disequilibrium<sup>8</sup> decays to  $r^2 = 0.2$  on average within 10 base pairs on autosomes and 30 base pairs on the X chromosome (Fig. 1c and Supplementary Fig. 1). This difference is expected because the population size of the X chromosome is three quarters that of autosomes, and the X chromosome can experience greater purifying selection because of exposure of deleterious recessive alleles in hemizygous males. There is little evidence of global population structure in the DGRP (Fig. 1d and Supplementary Fig. 2). The rapid decline in linkage disequilibrium locally and lack of global population structure are favourable for genome-wide association mapping.

Not all SNPs are fixed within individual DGRP lines (Supplementary Table 4). The expected inbreeding coefficient ( $F$ ) after 20 generations of full-sibling inbreeding<sup>1</sup> is  $F = 0.986$ ; therefore, we expect some SNPs to remain segregating by chance. Segregating SNPs can also arise from new mutations, or if natural selection opposes inbreeding, due to true overdominance for fitness at individual loci or associative overdominance due to complementary deleterious alleles that are closely linked or in segregating inversions.

We identified 390,873 microsatellite loci, 105,799 of which were polymorphic (Supplementary Table 5); 36,810 transposable element insertion sites and 197,402 total insertions (Supplementary Table 6). On average, each line contained 1,175 transposable element insertions (Supplementary Table 6), although most transposable element insertion sites (25,562) were present in only one line (Supplementary

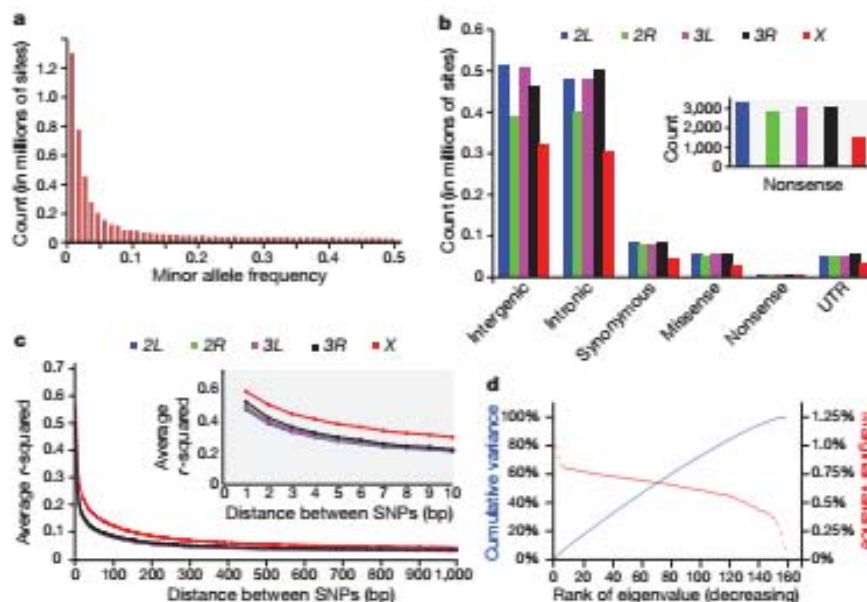
Table 7). We identified 149 transposable element families. The number of copies per family varied greatly from an average of 315.7 *INE-1* elements per line to an average of 0.003 *Gandalf-Dkoe-like* elements per line (Supplementary Table 8).

*Wolbachia pipiensis* is a maternally inherited bacterium found in insects, including *Drosophila*, and can affect reproduction<sup>9</sup>. We assessed *Wolbachia* infection status in the DGRP lines to account for it in analyses of genotype–phenotype associations, and found 51.2% of lines harbouring sufficient *Wolbachia* DNA to imply infection (Supplementary Table 9).

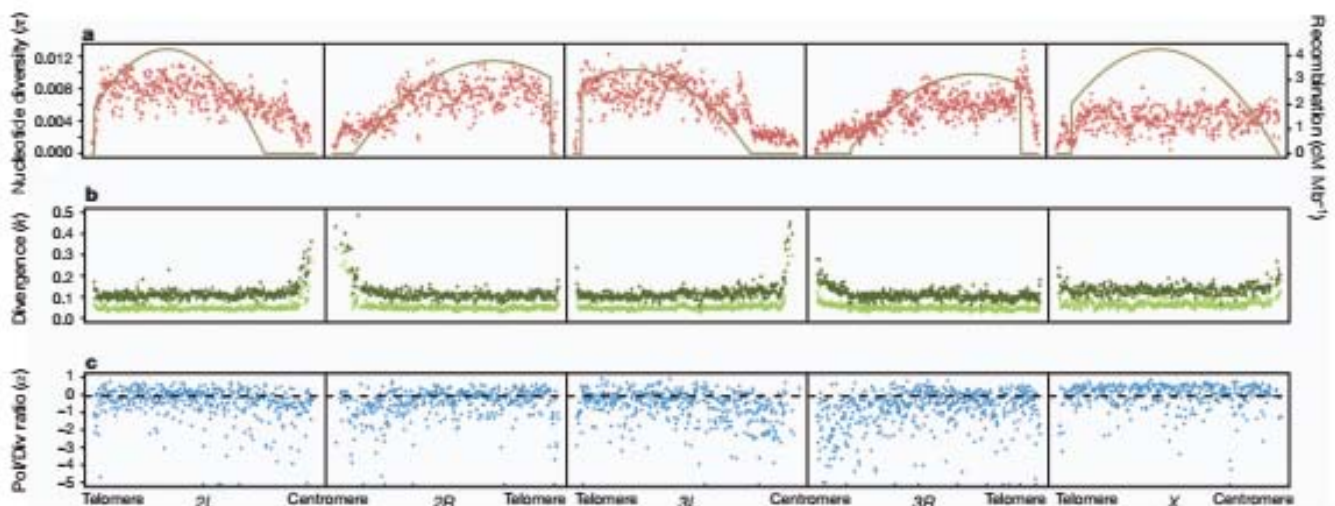
## Polymorphism and divergence

We used the DGRP Illumina sequence data and genome sequences from *Drosophila simulans* and *Drosophila yakuba*<sup>10</sup> to perform genome-wide analyses of polymorphism and divergence, assess the association of these parameters with genomic features and the recombination landscape, and infer the historical action of selection on a much larger scale than had been possible previously<sup>11–16</sup>. We computed polymorphism ( $\pi$  and  $\theta$ , refs 17 and 18) and divergence ( $k$ , ref. 19) for the whole genome, by chromosome arm (X, 2L, 2R, 3L, 3R), by chromosome region (three regions of equal size in Mb—telomeric, middle and centromeric), in 50-kbp non-overlapping windows, and by site class (synonymous and non-synonymous sites within coding sequences, and intronic, untranslated region (UTR) and intergenic sites) (Supplementary Tables 10 and 11).

Averaged over the entire genome,  $\pi = 0.0056$  and  $\theta = 0.0067$ , similar to previous estimates from North American populations<sup>16,20</sup>. Average polymorphism on the X chromosome ( $\pi_X = 0.0040$ ) is reduced relative to the autosomes ( $\pi_A = 0.0060$ ) ( $X/A$  ratio = 0.67, Wilcoxon test  $P = 0$ ), even after correcting for the  $X/A$  effective population size ( $X_{A_{eff}} = 0.0054$ , Wilcoxon test  $P < 0.00002$ ; Supplementary Table 10). Autosomal nucleotide diversity is reduced on average 2.4-fold in centromeric regions relative to non-centromeric regions, and at the telomeres (Fig. 2a and Supplementary Table 10), whereas diversity is relatively constant along the X chromosome. Thus,  $\pi_X > \pi_A$  in centromeric regions, but  $\pi_A > \pi_X$  in other chromosomal regions (Fig. 2a and Supplementary Table 10).



**Figure 1** | SNP variation in the DGRP lines. **a**, Site frequency spectrum. **b**, Numbers of SNPs per site class. **c**, Decay of linkage disequilibrium ( $r^2$ ) with physical distance for the five major chromosome arms. **d**, Lack of population structure. The red curve depicts the ranked eigenvalues of the genetic covariance matrix in decreasing order with respect to the marginal variance explained; the blue curve shows their cumulative sum as a fraction of the total with respect to cumulative variance explained. The partitioning of total genetic variance is balanced among the eigenvectors. The principal eigenvector explains < 1.1% of the total genetic variance.



**Figure 2** | Pattern of polymorphism, divergence,  $\alpha$  and recombination rate along chromosome arms in non-overlapping 50-kbp windows. **a**, Nucleotide polymorphism ( $\pi$ ). The solid curves give the recombination rate ( $\text{cM Mb}^{-1}$ ). **b**, Divergence ( $k$ ) for *D. simulans* (light green) and *D. yakuba* (dark green). **c**, Polymorphism to divergence ratio (Pol/Div), estimated as  $1 - [(\pi_{0-600}/\pi_{4-600})/(k_{0-600}/k_{4-600})]$ . An excess of 0-fold divergence relative to polymorphism ( $k_{0-600}/k_{4-600} > (\pi_{0-600}/\pi_{4-600})$ ) is interpreted as adaptive fixation whereas an excess of 0-fold polymorphism relative to divergence ( $\pi_{0-600}/\pi_{4-600} > (k_{0-600}/k_{4-600})$ ) indicates that weakly deleterious or nearly neutral mutations are segregating in the population.

Genes on the X chromosome evolve faster ( $k_X = 0.140$ ) than autosomal genes ( $k_A = 0.126$ ) ( $X/A$  ratio = 1.131, Wilcoxon test  $P = 0$ ) (Fig. 2b and Supplementary Table 10). Divergence is more uniform (coefficient of variation  $(CV)_k = 0.2841$ ) across chromosome arms than is polymorphism ( $(CV)_\pi = 0.4265$ ). The peaks of divergence near the centromeres could be attributable to the reduced quality of alignments in these regions. Patterns of divergence are similar regardless of the outgroup species used (Fig. 2b and Supplementary Table 11).

The pattern of polymorphism and divergence by site class is consistent within and among chromosomes ( $\pi_{k_{synonymous}} > \pi_{k_{missense}} > \pi_{k_{nonsynonymous}} > \pi_{k_{UTR}} > \pi_{k_{non-synonymous}}$ ), in agreement with previous studies on smaller data sets<sup>12,25</sup> (Supplementary Figs 3 and 4 and Supplementary Table 11). Polymorphism levels between synonymous and non-synonymous sites differ by an order of magnitude. Variation and divergence patterns within the site classes generally follow the same patterns observed overall, with reduced polymorphism for all site classes on the X chromosome relative to autosomes, increased X chromosome divergence relative to autosomes for all but synonymous sites, decreased polymorphism in centromeric regions, and greater variation among regions and arms for polymorphism than for divergence. Other diversity measures and more detailed patterns at different window-sizes for each chromosome arm can be accessed from the Population *Drosophila* Browser (popDrowser) (Table 1 and Methods).

### Recombination landscape

Evolutionary models of hitchhiking and background selection<sup>21,22</sup> predict a positive correlation between polymorphism and recombination rate. This expectation is realized in regions where recombination is less than  $2 \text{ cM Mb}^{-1}$  (Spearman's  $\rho = 0.471$ ,  $P = 0$ ), but recombination and polymorphism are independent in regions where recombination exceeds  $2 \text{ cM Mb}^{-1}$  (Spearman's  $\rho = -0.0044$ ,  $P = 0.987$ ) (Fig. 2a and Supplementary Table 12). The average rate of recombination of the X chromosome ( $2.9 \text{ cM Mb}^{-1}$ ) is greater than that of autosomes ( $2.1 \text{ cM Mb}^{-1}$ ), which may account for the low overall X-linked correlation between recombination rate and  $\pi$ . The lack of correlation between recombination and divergence (Supplementary Table 12) excludes mutation associated with recombination as the cause of the correlation. We assessed the independent effects of recombination rate, divergence, chromosome region and gene density on nucleotide variation of autosomes and the X chromosome (Supplementary Table 13). Recombination is the major predictor of

polymorphism on the X chromosome and autosomes; however, the significant effect of autosomal chromosome region remains after accounting for variation in recombination rates between centromeric and non-centromeric regions.

### Selection regimes

We used the standard<sup>23</sup> and generalized<sup>12,24,25</sup> McDonald Kreitman tests (MKT) to scan the genome for evidence of selection. These tests

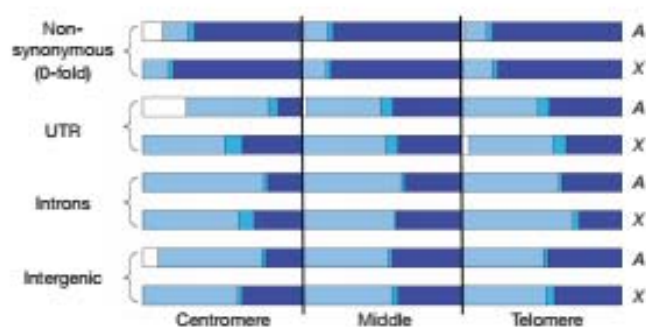
**Table 1** | Community resources

Resource	Location
DGRP lines	Bloomington <i>Drosophila</i> Stock Center <a href="http://flystocks.bio.indiana.edu/Browse/RAL.php">http://flystocks.bio.indiana.edu/Browse/RAL.php</a>
Sequences	Baylor College of Medicine Human Genome Sequencing Center <a href="http://www.hgsc.bcm.tmc.edu/project-species-i-DGRP_lines.hgsc">http://www.hgsc.bcm.tmc.edu/project-species-i-DGRP_lines.hgsc</a>
Read alignments	National Center for Biotechnology Information Short Read Archive <a href="http://www.ncbi.nlm.nih.gov/sra?term=DGRP">http://www.ncbi.nlm.nih.gov/sra?term=DGRP</a> Mackay Laboratory <a href="http://dgrp.gnets.ncsu.edu/">http://dgrp.gnets.ncsu.edu/</a>
SNPs	Baylor College of Medicine Human Genome Sequencing Center <a href="http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze1_July_2010/snp_calls/">http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze1_July_2010/snp_calls/</a> National Center for Biotechnology Information dbSNP <a href="http://www.ncbi.nlm.nih.gov/SNP/snp_viewBatch.cgi?sbid=1052186">http://www.ncbi.nlm.nih.gov/SNP/snp_viewBatch.cgi?sbid=1052186</a> Mackay Laboratory <a href="http://dgrp.gnets.ncsu.edu/">http://dgrp.gnets.ncsu.edu/</a>
Microsatellites	Baylor College of Medicine Human Genome Sequencing Center <a href="http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze1_July_2010/microsat_calls/">http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze1_July_2010/microsat_calls/</a> Mittelman Laboratory <a href="http://genome.vbi.vt.edu/public/DGRP/">http://genome.vbi.vt.edu/public/DGRP/</a>
Transposable elements	Mackay Laboratory <a href="http://dgrp.gnets.ncsu.edu/">http://dgrp.gnets.ncsu.edu/</a>
Molecular population genomics	PopDrowser <a href="http://popdrowser.uab.cat">http://popdrowser.uab.cat</a>
Phenotypes	Mackay Laboratory <a href="http://dgrp.gnets.ncsu.edu/">http://dgrp.gnets.ncsu.edu/</a>
Genome-wide association analysis	Mackay Laboratory <a href="http://dgrp.gnets.ncsu.edu/">http://dgrp.gnets.ncsu.edu/</a>

compare the ratio of polymorphism at a selected site with that of a neutral site to the ratio of divergence at a selected site to divergence at a neutral site. The standard MKT is applied to coding sequences, and synonymous and non-synonymous sites are used as putative neutral and selected sites, respectively. The generalized MKT is applied to non-coding sequences and uses fourfold degenerate sites as neutral sites. Using polymorphism and divergence data avoids confounding inference of selection with mutation rate differences, and restricting the tests to closely linked sites controls for shared evolutionary history<sup>26–28</sup>. We infer adaptive divergence when there is an excess of divergence relative to polymorphism, and segregation of slightly deleterious mutations when there is an excess of polymorphism over divergence. Estimates of  $\alpha$ , the proportion of adaptive divergence, are biased downwards by low frequency, slightly deleterious mutations<sup>29,30</sup>. Rather than eliminate low frequency variants<sup>31</sup>, we incorporated information on the site frequency distribution in the MKT test framework to obtain estimates of the proportion of sites that are strongly deleterious ( $d$ ), weakly deleterious ( $b$ ), neutral ( $f$ ) and recently neutral ( $\gamma$ ) at segregating sites, as well as unbiased estimates of  $\alpha$  (Supplementary Methods).

### Deleterious and neutral sites

Averaged over the entire genome, we infer that 58.5% of the segregating sites are neutral or nearly neutral, 1.9% are weakly deleterious and 39.6% are strongly deleterious. However, these proportions vary between the *X* chromosome and autosomes, site classes and chromosome regions (Supplementary Tables 14–16 and Fig. 3). Non-synonymous sites are the most constrained ( $d = 77.6\%$ ), whereas in non-coding sites  $d$  ranges from 29.1% in 5' UTRs to 41.3% in 3' intergenic regions. The inferred pattern of selection differs between autosomal centromeric and non-centromeric regions:  $d$  is reduced and  $f$  is increased in centromeric regions for all site categories (Fig. 3). We observe an excess of polymorphism relative to divergence in autosomal centromeric regions, even after correcting for weakly deleterious mutations, implying a relaxation of selection from the time of separation of *D. melanogaster* and *D. yakuba*. Because selection coefficients depend on the effective population size<sup>32</sup> ( $N_e$ ), this could occur if the recombination rate has specifically diminished in centromeric regions during the divergence between *D. melanogaster* and *D. yakuba*; or with an overall reduction of  $N_e$  associated with the colonization of North American habitats<sup>33,34</sup>. In the latter case, we expect a genome-wide signature of an excess of low-frequency polymorphisms and of polymorphism relative to divergence, exacerbated in regions of low recombination. We indeed find an excess of low-frequency polymorphism relative to neutral expectation as indicated by the negative estimates of Tajima's  $D$  statistic<sup>35</sup>



**Figure 3** | The fraction of alleles segregating under different selection regimes by site class and chromosome region, for the autosomes (A) and the *X* chromosome (X). The selection regimes are strongly deleterious ( $d$ , dark blue), weakly deleterious ( $b$ , blue), recently neutral ( $\gamma$ , white) and old neutral ( $f - \gamma$ , light blue). Each chromosome arm has been divided in three regions of equal size (in Mb): centromere, middle and telomere.

( $D = -0.686$  averaged over the whole genome and  $D = -0.997$  in autosomal centromeric regions). In contrast, the *X* chromosome does not show a differential pattern of selection in the centromeric region, has a lower fraction of relaxation of selection, fewer neutral alleles, and a higher percentage of strongly deleterious alleles for all site classes and regions (Fig. 3 and Supplementary Tables 14–16).

Transposable element insertions are thought to be largely deleterious. There are more singleton insertions in regions of high recombination ( $\geq 2$  cM Mb<sup>-1</sup>) and more insertions shared in multiple lines in regions of low recombination ( $< 2$  cM Mb<sup>-1</sup>) (Fisher's exact test  $P = 0$ ), and comparison of observed and expected site occupancy spectra reveals an excess of singleton insertions ( $P = 0$ , Supplementary Fig. 5).

### Adaptive fixation

We find substantial evidence for positive selection in autosomal non-centromeric regions and the *X* chromosome (Fig. 2c and Supplementary Tables 15 and 17). We estimated  $\alpha$  by aggregating all sites in each region analysed to avoid underestimation by averaging across genes<sup>36</sup> in comparisons of chromosomes, regions and site classes. We also computed the direction of selection, DoS<sup>37</sup>, which is positive with adaptive selection, zero under neutrality and negative when weakly deleterious or new nearly neutral mutations are segregating. Estimates of  $\alpha$  from the standard and generalized MKT indicate that on average 25.2% of the fixed sites between *D. melanogaster* and *D. yakuba* are adaptive, ranging from 30% in introns to 7% in UTR sites (Supplementary Fig. 6). Estimates of DoS and  $\alpha$  are negative for non-synonymous and UTR sites in the autosomal centromeres, consistent with underestimating the fraction of adaptive substitutions in regions of low recombination because weakly deleterious or nearly neutral mutations are more common than adaptive fixations. The majority of adaptive fixation on autosomes occurs in non-centromeric regions (Fig. 2c). We find over four times as many adaptive fixations on the *X* chromosome relative to autosomes. The pattern holds for all site classes, in particular non-synonymous sites and UTRs, as well as individual genes, and is not solely due to the autosomal centromeric effect (Supplementary Table 15 and Supplementary Figs 6 and 7). Finally, when we consider DoS in recombination environments above and below 2 cM Mb<sup>-1</sup>, we find greater adaptive propensity in genes whose recombination context is  $\geq 2$  cM Mb<sup>-1</sup> (Wilcoxon test,  $P = 0$ ; Supplementary Fig. 8).

To understand the global patterns of divergence and constraint across functional classes of genes, we examined the distributions of  $\omega$  ( $d_N/d_S$ , the ratio of non-synonymous to synonymous divergence) and DoS across gene ontology (GO) categories. The 4.9% GO categories with significantly elevated DoS include the biological process categories of behaviour, developmental process involved in reproduction, reproduction and ion transport (Supplementary Table 18). Recombination context is the major determinant of variation in DoS (Supplementary Table 19) whereas GO category is as important as recombinational context for predicting variation in  $\omega$  (Supplementary Table 19).

GO categories enriched for positive DoS values differ from those associated with high values of  $\omega$  (Supplementary Table 18), indicating that positive selection does not occur necessarily on genes with high  $\omega$  values. If adaptive substitutions are common, high values of  $\omega$  reflect the joint contributions of neutral and adaptive substitutions. Further, equating high constraint (low  $\omega$ ) with functional importance overlooks the functional role of adaptive changes<sup>35</sup>. Unlike  $\omega$ , DoS takes into account the constraints inferred from the current polymorphism, distinguishing negative, neutral and adaptive selection.

### Genome-wide genotype-phenotype associations

We measured resistance to starvation stress, chill coma recovery time and startle response<sup>38</sup> in the DGRP. We found considerable genetic variation for all traits, with high broad sense heritabilities. We also found variation in sex dimorphism for starvation resistance and chill

coma recovery with cross-sex genetic correlations significantly different from unity (Supplementary Tables 20–22).

We performed genome-wide association analyses for these traits, using the 2,490,165 SNPs and 77,756 microsatellites for which the minor allele was represented in four or more lines, using single-locus analyses pooled across sexes and separately for males and females. At  $P < 10^{-5}$  ( $P < 10^{-6}$ ), we find 203 (32) SNPs and 2 (0) microsatellites associated with starvation resistance; 90 (7) SNPs and 4 (2) microsatellites associated with startle response; and 235 (45) SNPs and 5 (3) microsatellites associated with chill coma recovery time (Fig. 4a, Supplementary Fig. 9 and Supplementary Tables 23 and 24). The minor allele frequencies for most of the associated SNPs are low, and there is an inverse relationship between effect sizes and minor allele frequency (Supplementary Fig. 10).

The DGRP is a powerful tool for rapidly reducing the search space for molecular variants affecting quantitative traits from the entire genome to candidate polymorphisms and genes. Although we cannot infer which of these polymorphisms are causal due to linkage disequilibrium between SNPs in close physical proximity as well as occasional spurious long range linkage disequilibrium (Fig. 4a and Supplementary Fig. 9), the candidate gene lists are likely to be enriched for causal variants. The majority of associations are in computationally predicted genes or genes with annotated functions not obviously associated with the three traits. However, genes previously associated with startle response<sup>29</sup> (*Sema-1a* and *Eip75B*) and starvation resistance<sup>40</sup> (*pnt*) were identified in this study; and a SNP in *CG3213*, previously identified in a *Drosophila* obesity screen<sup>41</sup>, is associated with variation in starvation resistance. Several genes associated with quantitative traits are rapidly evolving (*psq*, *Egfr*; Supplementary Tables 17 and 23) or are plausible candidates based on SNP or gene ontology annotations (Supplementary Table 23).

### Predicting phenotypes from genotypes

We used regression models to predict trait phenotypes from SNP genotypes and estimate the total variance explained by SNPs. The latter cannot be done by summing the individual contributions of the single marker effects because markers are not completely independent, and estimates of effects of single markers are biased when more than one locus affecting the trait segregates in the population. We derived gene-centred multiple regression models to estimate the effects of multiple SNPs simultaneously. In all cases 6–10 SNPs explain from 51–72% of the phenotypic variance and 65–90% of the genetic variance (Supplementary Tables 25 and 26 and Supplementary Figs 11–13). We also derived partial least square regression models using all SNPs for which the single marker effect was significant

at  $P < 10^{-5}$ . These models explain 72–85% of the phenotypic variance (Fig. 4b, c and Supplementary Fig. 14).

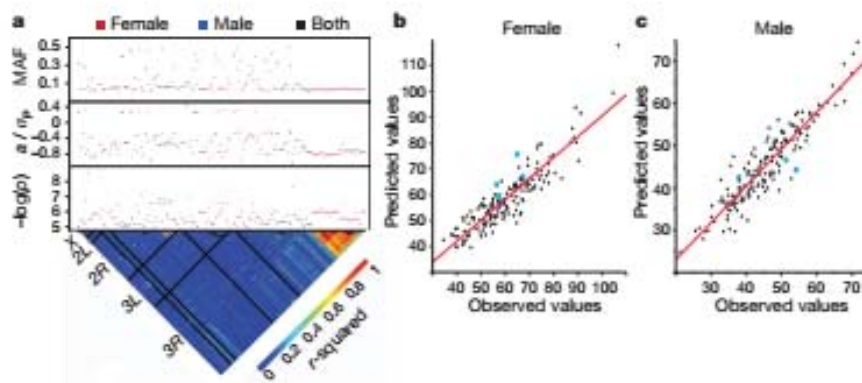
### Discussion

The DGRP lines, sequences, variant calls, phenotypes and web tools for molecular population genomics and genome-wide association analysis are publicly available (Table 1). The DGRP lines contain at least 4,672,297 SNPs, 105,799 polymorphic microsatellites and 36,810 transposable elements, as well as insertion/deletion events and copy number variants and are a valuable resource for understanding the genetic architecture of quantitative traits of ecological and evolutionary relevance as well as *Drosophila* models of human quantitative traits. These novel mutations have survived the sieve of natural selection and will enhance the functional annotation of the *Drosophila* genome, complementing the *Drosophila* Gene Disruption Project<sup>42</sup> and the *Drosophila* modENCODE project<sup>43</sup>.

Genome-wide molecular population genetic analyses show that patterns of polymorphism, but not divergence, differ by autosomal chromosome region, and between the X chromosome and autosomes. Polymorphism is lower in autosomal centromeric than non-centromeric regions, but not for the X chromosome. We propose that the correlation of polymorphism with recombination in regions where recombination is  $< 2 \text{ cM Mb}^{-1}$  is due to the reduced effective population size in regions of low recombination<sup>8</sup>. Selection is less efficient in regions of low recombination<sup>32</sup>, consistent with our observation that the fraction of strongly deleterious mutations and positively selected sites are reduced in these regions.

All molecular population genomic analyses support the ‘faster X’ hypothesis<sup>44</sup>. Relative to the autosomes, the X chromosome shows lower polymorphism, faster rates of molecular evolution, a higher percentage of gene regions undergoing adaptive evolution, a higher fraction of strongly deleterious sites, and a lower level of weak negative selection and relaxation of selection. New X-linked mutations are directly exposed to selection each generation in hemizygous males, and the X chromosome has greater recombination than autosomes<sup>44</sup>; both of these factors could contribute to this observation.

Genome-wide association analyses of three fitness-related quantitative traits reveal hundreds of novel candidate genes, highlighting our ignorance of the genetic basis of complex traits. Most variants associated with the traits are at low frequency, and there is an inverse relationship between frequency and effect. Given that low-frequency alleles are likely to be deleterious for traits under directional or stabilizing selection, these results are consistent with the mutation–selection balance hypothesis<sup>1</sup> for the maintenance of quantitative genetic variation. Regression models incorporating significant SNPs



**Figure 4 | Genotype–phenotype associations for starvation resistance.** a, Genome-wide association results for significant SNPs. The lower triangle depicts linkage disequilibrium ( $r^2$ ) among SNPs, with the five major chromosome arms demarcated by black lines. The upper panels give the significance threshold ( $-\log(p)$ , uncorrected for multiple tests), the effect in phenotypic standard deviation units, and the minor allele frequency (MAF). b, c, Partial least squares regressions of phenotypes predicted using SNP data on observed phenotypes. The blue dots represent the predicted and observed phenotypes of lines that were not included in the initial study. b, Females ( $r^2 = 0.81$ ); c, males ( $r^2 = 0.85$ ).



explain most of the phenotypic variance of the traits, in contrast with human association studies, where significant SNPs have tiny effects and together explain a small fraction of the total phenotypic variance<sup>7</sup>. If the genetic architecture of human complex traits is also dominated by low-frequency causal alleles, we expect estimates of effect size based on linkage disequilibrium with common variants to be strongly biased downwards.

In the future, the full power of *Drosophila* genetics can be applied to validating marker-trait associations: mutations, RNA interference constructs and quantitative trait loci mapping populations. The DGRP is an ideal resource for systems genetics analyses of the relationship between molecular variation, causal molecular networks and genetic variation for complex traits<sup>4,38,45</sup>, and will anchor evolutionary studies in comparison with sequenced *Drosophila* species to assess to what extent variation within a species corresponds to variation among species.

## METHODS SUMMARY

The full Methods are in the Supplementary Information. Information on sequencing and bioinformatics includes methods for DNA isolation; library construction and genomic sequencing; sequence read alignment; SNP, microsatellite and transposable element identification; genotypes for assurance of sample identity; and *Wolbachia* detection. Methods for molecular population genomics analysis include details of recombination estimates; diversity measures, linkage disequilibrium and neutrality tests; software used for population genomic analysis; data visualization (popDrowser); standard and generalized McDonald-Kreitman tests, statistical analysis methods; quality assessment and data filtering; and gene ontology analyses. Methods for quantitative genetic analyses include phenotype measures, quantitative genetic analyses of phenotypes, statistical analyses of genotype-phenotype associations and predictive models, and a web-based association analysis pipeline.

Received 13 July; accepted 21 December 2011.

- Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics* 4th edn (Longman, 1996).
- Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, 1998).
- Flint, J. & Mackay, T. F. C. Genetic architecture of quantitative traits in flies, mice and humans. *Genome Res.* **19**, 723–733 (2009).
- Mackay, T. F. C., Stone, E. A. & Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. *Nature Rev. Genet.* **10**, 565–577 (2009).
- Alshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
- Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728–731 (2008).
- Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
- Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
- Werren, J. H. Biology of *Wolbachia*. *Annu. Rev. Entomol.* **42**, 587–609 (1997).
- Clark, A. G. et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218 (2007).
- Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024 (2002).
- Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**, 1149–1152 (2005).
- Presgraves, D. C. Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr. Biol.* **15**, 1651–1656 (2005).
- Casillas, S., Barbadilla, A. & Bergman, C. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Mol. Biol. Evol.* **24**, 2222–2234 (2007).
- Sella, G. et al. Pervasive natural selection in the *Drosophila* genome? *PLoS Genet.* **5**, e1000495 (2009).
- Sackton, T. B. et al. Population genomic inferences from sparse high-throughput sequencing of two populations of *Drosophila melanogaster*. *Genome Biol. Evol.* **1**, 449–465 (2009).
- Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, 1987).
- Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
- Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* vol. 3 (eds Munro, H. N. & Allison, J. B.) 21–132 (Academic Press, 1969).
- Andolfatto, P. & Przeworski, M. Regions of lower crossing over harbor more rare variants in African *Drosophila melanogaster*. *Genetics* **158**, 657–665 (2001).

- Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992).
- Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
- McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
- Jenkins, D. L., Orton, C. A. & Brookfield, J. F. A test for adaptive change in DNA sequences controlling transcription. *Proc. R. Soc. Lond. B* **261**, 203–207 (1995).
- Egea, R., Casillas, S. & Barbadilla, A. Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.* **36**, W157–W162 (2008).
- Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).
- Nielsen, R. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641–647 (2001).
- Eyre-Walker, A. Changing effective population size and the McDonald-Kreitman test. *Genetics* **162**, 2017–2024 (2002).
- Charlesworth, J. & Eyre-Walker, A. The McDonald-Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol.* **25**, 1007–1015 (2008).
- Eyre-Walker, A. & Keightley, P. D. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* **26**, 2097–2108 (2009).
- Fay, J. C., Wyckoff, G. J. & Wu, C. I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024–1026 (2002).
- Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).
- David, J. R. & Capi, P. Genetic variation of *Drosophila melanogaster* natural populations. *Trends Genet.* **4**, 106–111 (1988).
- Begun, D. J. & Aquadro, C. F. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature* **365**, 548–550 (1993).
- Tajima, F. Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
- Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024 (2002).
- Stoletzki, N. & Eyre-Walker, A. Estimation of the neutrality index. *Mol. Biol. Evol.* **28**, 63–70 (2011).
- Ayroles, J. F. et al. Systems genetics of complex traits in *Drosophila melanogaster*. *Nature Genet.* **41**, 299–307 (2009).
- Yamamoto, A. et al. Neurogenetic networks for startle-induced locomotion in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **105**, 12393–12398 (2008).
- Harbison, S. T., Yamamoto, A. H., Fanara, J. J., Norga, K. K. & Mackay, T. F. C. Quantitative trait loci affecting starvation resistance in *Drosophila melanogaster*. *Genetics* **166**, 1807–1823 (2004).
- Pospisilik, J. A. et al. *Drosophila* genome-wide obesity screen reveals hedgehog as a determinant of brown versus white adipose cell fate. *Cell* **140**, 148–160 (2010).
- Bellen, H. J. et al. The BDGP gene disruption project: single transposon insertions associated with 40% of *Drosophila* genes. *Genetics* **167**, 761–781 (2004).
- The ModENCODE Consortium. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
- Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130**, 113–146 (1987).
- Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm. Genome* **18**, 389–401 (2007).

Supplementary Information is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** This work was supported by National Institutes of Health grant GM 45146 to T.F.C.M., E.A.S. and R.R.H.A.; R01 GM 059469 to R.R.H.A., MCI BFU 2009-09504 to A.B., R01 GM 085183 to K.R.T., NIHRI U54 HG003273 to R.A.G.; and an award through the NVIDIA Foundation's "Compute the Cure" programme to D.M.

**Author Contributions** T.F.C.M., S.R. and R.A.G. conceived the project. T.F.C.M., S.R., A.B. and E.A.S. wrote the main manuscript. T.F.C.M., S.R., A.B., E.A.S., J.F.A., K.R.T., J.M.C., C.M.B. and D.M. wrote the Supplementary methods. M.M.M., C.B., K.P.B., M.A.C., L.C., L.D., Y.H., M.J., J.C.J., S.N.J., K.W.J., F. Lara, F. Lawrence, S.L.L., R.F.L., M.M., D.M.M., L.N., L.M., L.P., L.L.P., C.Q., J.G.R., S.M.R., L.T., K.C.W., Y.-Q.W., A.Y. and Y.Z. performed experiments. T.F.C.M., A.B., J.F.A., D.Z., S.C., M.M.M., J.M.C., M.F.R., M.B., D.C., R.S.L., A.M., C.M.B., K.R.T., D.M. and E.A.S. did the bioinformatics and data analysis. J.F.A., S.C., M.M.M., Z.H., P.L., M.R., J.R. and E.A.S. wrote the Methods and did the web site development. R.R.H.A. contributed resources.

**Author Information** Sequences have been deposited at the National Center for Biotechnology Information Short Read Archives (<http://www.ncbi.nlm.nih.gov/sra?term=DGRP>). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at [www.nature.com/nature](http://www.nature.com/nature). The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at [www.nature.com/nature](http://www.nature.com/nature). Correspondence and requests for materials should be addressed to T.F.C.M. ([trudy\\_mackay@ncsu.edu](mailto:trudy_mackay@ncsu.edu)).



---

## **Annex C**

Natural variation in genome architecture  
among 205 *Drosophila melanogaster* genetic  
reference panel lines

## Resource

# Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines

Wen Huang,<sup>1,10</sup> Andreas Massouras,<sup>2,3,10</sup> Yutaka Inoue,<sup>4</sup> Jason Peiffer,<sup>1</sup> Miquel Ràmia,<sup>5</sup> Aaron M. Tarone,<sup>6</sup> Lavanya Turlapati,<sup>1</sup> Thomas Zichner,<sup>7</sup> Dianhui Zhu,<sup>8,12</sup> Richard F. Lyman,<sup>1</sup> Michael M. Magwire,<sup>1,13</sup> Kerstin Blankenburg,<sup>8</sup> Mary Anna Carbone,<sup>1</sup> Kyle Chang,<sup>8</sup> Lisa L. Ellis,<sup>6</sup> Sonia Fernandez,<sup>8</sup> Yi Han,<sup>8</sup> Gareth Highnam,<sup>9</sup> Carl E. Hjelman,<sup>6</sup> John R. Jack,<sup>1</sup> Mehwish Javaid,<sup>8</sup> Joy Jayaseelan,<sup>8</sup> Divya Kalra,<sup>8</sup> Sandy Lee,<sup>8</sup> Lora Lewis,<sup>8</sup> Mala Munidasa,<sup>8</sup> Fiona Onger, <sup>8</sup> Shohba Patel,<sup>8</sup> Lora Perales,<sup>8</sup> Agapito Perez,<sup>8</sup> LingLing Pu,<sup>8</sup> Stephanie M. Rollmann,<sup>1,14</sup> Robert Ruth,<sup>8</sup> Nehad Saada,<sup>8</sup> Crystal Warner,<sup>8,15</sup> Aneisa Williams,<sup>8</sup> Yuan-Qing Wu,<sup>8</sup> Akihiko Yamamoto,<sup>1</sup> Yiqing Zhang,<sup>8</sup> Yiming Zhu,<sup>8</sup> Robert R.H. Anholt,<sup>1</sup> Jan O. Korbel,<sup>7</sup> David Mittelman,<sup>9</sup> Donna M. Muzny,<sup>8</sup> Richard A. Gibbs,<sup>8</sup> Antonio Barbadilla,<sup>5,11</sup> J. Spencer Johnston,<sup>6,11</sup> Eric A. Stone,<sup>1,11</sup> Stephen Richards,<sup>8,11</sup> Bart Deplancke,<sup>2,3,11</sup> and Trudy F.C. Mackay<sup>1,11,16</sup>

<sup>1–9</sup>[Author affiliations appear at the end of the paper.]

The *Drosophila melanogaster* Genetic Reference Panel (DGRP) is a community resource of 205 sequenced inbred lines, derived to improve our understanding of the effects of naturally occurring genetic variation on molecular and organismal phenotypes. We used an integrated genotyping strategy to identify 4,853,802 single nucleotide polymorphisms (SNPs) and 1,296,080 non-SNP variants. Our molecular population genomic analyses show higher deletion than insertion mutation rates and stronger purifying selection on deletions. Weaker selection on insertions than deletions is consistent with our observed distribution of genome size determined by flow cytometry, which is skewed toward larger genomes. Insertion/deletion and single nucleotide polymorphisms are positively correlated with each other and with local recombination, suggesting that their nonrandom distributions are due to hitchhiking and background selection. Our cytogenetic analysis identified 16 polymorphic inversions in the DGRP. Common inverted and standard karyotypes are genetically divergent and account for most of the variation in relatedness among the DGRP lines. Intriguingly, variation in genome size and many quantitative traits are significantly associated with inversions. Approximately 50% of the DGRP lines are infected with *Wolbachia*, and four lines have germline insertions of *Wolbachia* sequences, but effects of *Wolbachia* infection on quantitative traits are rarely significant. The DGRP complements ongoing efforts to functionally annotate the *Drosophila* genome. Indeed, 15% of all *D. melanogaster* genes segregate for potentially damaged proteins in the DGRP, and genome-wide analyses of quantitative traits identify novel candidate genes. The DGRP lines, sequence data, genotypes, quality scores, phenotypes, and analysis and visualization tools are publicly available.

[Supplemental material is available for this article.]

Studies in *Drosophila melanogaster* have revealed basic principles and mechanisms underlying fundamental genetic concepts of linkage and recombination and were instrumental in identifying canonical and evolutionarily conserved cell signaling pathways.

#### <sup>10</sup>Joint first authors

#### <sup>11</sup>Senior authors

Present addresses: <sup>12</sup>Chevron Inc., Houston, TX 77002, USA; <sup>13</sup>Syngenta, Research Triangle Park, NC 27709, USA; <sup>14</sup>Department of Biological Sciences, University of Cincinnati, Cincinnati, OH 45221, USA; <sup>15</sup>Shell International Exploration and Production, Inc., Houston, TX 77082-3101, USA.

#### <sup>16</sup>Corresponding author

E-mail [trudy\\_mackay@ncsu.edu](mailto:trudy_mackay@ncsu.edu)

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.171546.113>. Freely available online through the Genome Research Open Access option.

Most *D. melanogaster* genes are evolutionarily conserved, leading to fly models for understanding common human diseases and behavioral disorders, dipteran disease vectors, and insects impacting agriculture, medicine, and forensics. Despite nearly a century of research on *D. melanogaster*, however, a large fraction of its coding and noncoding sequence has no known function (McQuilton et al. 2012). Recent efforts to induce mutations in every protein coding gene utilize transposable elements (Bellen et al. 2004, 2011), which have a different spectrum of allelic effects than SNPs and small insertions and deletions (indels). Comprehensive efforts to identify regulatory DNA elements in *Drosophila* (The

© 2014 Huang et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0>.

modENCODE Consortium et al. 2010) have attributed functional effects to noncoding DNA, further complicating efforts to dissect the genotype-phenotype map. In addition, the vast majority of genetic analyses in *D. melanogaster* have used a few "wild type" strains representing a tiny sample of genetic diversity. Mutational effects in one genetic background are often enhanced or suppressed in other backgrounds (Mackay 2014). Such epistatic interactions provide a window for visualizing genetic interaction networks. In addition, *D. melanogaster* has a rich history as a model organism for population and quantitative genetics, generating inferences about regions under purifying natural selection independent of functional analyses and highlighting the contribution of common and rare variants in protein coding as well as regulatory sequences to the genetic architecture of complex traits (Flint and Mackay 2009; Mackay et al. 2009).

Efforts to utilize naturally occurring genetic variation in *D. melanogaster* to add to our understanding of functional DNA elements have been greatly expedited by the *Drosophila* Genetic Reference Panel (DGRP), a publicly available population of 205 sequenced inbred lines. Previously, we cataloged SNPs segregating in 168 DGRP lines (DGRP Freeze 1.0) (Mackay et al. 2012) and non-SNP variants in a subset of 39 lines (Massouras et al. 2012; Zichner et al. 2013). Here, we report the DGRP Freeze 2.0 with sequences of all lines and genotypes for SNP and non-SNP variants (indels, tandem duplications, and complex variants). We describe cytogenetic analysis of inversions, *Wolbachia* infection status, variation in genome size, molecular population genetics of indels and inversions, functional analyses of segregating variants, and online tools for association mapping of complex traits.

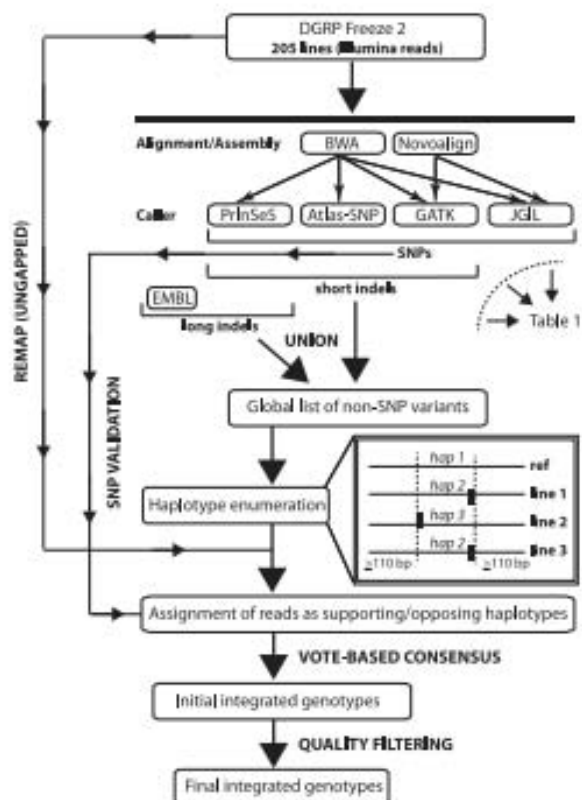
## Results

### Catalog of molecular polymorphism in the DGRP

We obtained Illumina sequences for 48 DGRP lines that were not sequenced previously or for which only 454 sequence data were available, as well as for six DGRP lines with low Freeze 1.0 coverage (Supplemental Data File S1). We aligned sequence reads to the *D. melanogaster* genome using BWA (Li and Durbin 2010) and Novoalign (Novocraft.com), recalibrated base quality scores, and locally realigned reads. The average mapped sequence coverage was 27× per line (Supplemental Data File S1).

There are many algorithms for detecting SNP and non-SNP variants from short-read sequence data (Massouras et al. 2010; McKenna et al. 2010; Medvedev et al. 2010; Shen et al. 2010; Alkan et al. 2011; Rausch et al. 2012; Stone 2012). Identification of non-SNP variants is challenging with short reads (Onishi-Seebacher and Korbel 2011), since structural variants can produce alternative alignments and variant calls for the same variant. Methods combining several approaches to generate a consensus variant list give a lower false positive rate than individual methods (Mills et al. 2011; Zichner et al. 2013). Variant call quality can be further enhanced by genotyping to test if variants in the population are also present in the line under consideration (Waszak et al. 2010; Handsaker et al. 2011). In regions of low read depth, such genotyping may be possible even though variants cannot be discovered de novo. In this study, we used seven variant callers and integrated genotyping (Fig. 1) to comprehensively map genomic variation in 205 DGRP lines.

On average, the methods called over 580,000 SNPs, and 130,000 small (<100 bp) and 1400 large (≥100 bp) non-SNP variants per line; however, there was heterogeneity in the number of variants called by each method and the overall concordance



**Figure 1.** Flowchart of the integrated genotyping procedure used to call SNP and non-SNP variants. Seven different variant calling methods were used to derive a consensus list of variant calls. The variant calls were grouped into haplotype bins (indicated by dashed vertical lines) such that there is a region on both sides of each region containing two or more regions of at least 110 bp with no non-SNP variants in any line. The variable regions and their 110-bp flanking regions were used to derive the sequences of alternative haplotypes against which reads are aligned. Finally, reads were aligned and genotypes called, followed by quality filtering that accounted for the experimental design.

among methods (Table 1). Therefore, we implemented an integrated genotyping algorithm, first using the combined data from all variant calling methods to update the genotypes of each DGRP line, then using the 205 variant call lists to genotype each DGRP line (Fig. 1). We identified 6,149,882 unique variants, including 4,853,802 SNPs and 1,296,080 non-SNP variants. The majority (98.28%) of the non-SNP variants were <100 bp.

### Validation of genotype calls

We used three strategies to validate genotype calls. First, we targeted 384 small (1–18 bp) indels affecting coding regions and 384 randomly chosen larger (30–313 bp) indels for validation by Sanger sequencing in five DGRP lines. A total of 315 small and 384 large indels were successfully assayed with Sanger technology for at least three lines. Of the 1463 small indel/line and 1876 large indel/line combinations with both Sanger and Illumina calls, 1458 (99.66%) and 1872 (99.79%), respectively, were concordant (Supplemental Data Files S2, S3).

Second, we performed high-density tiling microarray-based validation experiments using published data for six DGRP lines

**Table 1.** Comparison of genotyping methods for (A) SNPs, (B) short (<100 bp) non-SNP variants, (C) long (≥100-bp non-SNP variants)

(A)	PrinSeS/BWA	GATK/Novoalign	GATK/BWA	Atlas-SNP/BWA	JGIL/BWA	JGIL/Novoalign
PrinSeS/BWA	<b>635,828</b>	557,694	582,538	435,257	556,629	541,012
GATK/Novoalign	84%	<b>583,225</b>	569,871	443,982	538,989	538,426
GATK/BWA	86%	89%	<b>627,295</b>	449,293	571,782	548,159
Atlas-SNP/BWA	66%	74%	71%	<b>459,224</b>	425,312	419,496
JGIL/BWA	81%	83%	86%	66%	<b>606,778</b>	557,706
JGIL/Novoalign	81%	87%	84%	68%	89%	<b>576,940</b>
(B)	PrinSeS/BWA	GATK/Novoalign	GATK/BWA	Atlas-SNP/BWA		
PrinSeS/BWA	<b>174,550</b>	102,531	106,969	81,912		
GATK/Novoalign	55%	<b>115,562</b>	97,154	75,415		
GATK/BWA	54%	65%	<b>131,554</b>	82,850		
Atlas-SNP/BWA	41%	51%	53%	<b>106,887</b>		
(C)	PrinSeS/BWA	EMBL				
PrinSeS/BWA	<b>1672</b>	399				
EMBL	19%	<b>1138</b>				

The numbers on the diagonal (boldface) are the average numbers of variants called per line by each method. The numbers above the diagonal are the average numbers of variants found in common between the methods indicated by the row and column labels. The numbers below the diagonal are the percentage of calls that agree between the indicated pair of methods for DGRP sites at which both methods identify at least one non-reference base.

(Zichner et al. 2013) to assess the accuracy of the genotyping of larger deletions (>25 bp). We evaluated 3930 deletions ranging in size from 27 to 7533 bp. Of 5957 deletion/line comparisons, 5170 (86.8%) were true positives and 787 (13.2%) were false positives (Supplemental Fig. S1).

Third, we used the 454 sequence data from 38 lines (Mackay et al. 2012; Supplemental Table S1) to validate SNP and non-SNP calls. We used our integrated genotyping algorithm to call variants but restricted the input variant list to the final calls from the Illumina genotyping analysis. Using the same genotyping pipeline but a different sequencing chemistry serves to validate the Illumina data generation process. We used Fisher's exact test to statistically evaluate whether the Illumina and 454 genotypes were concordant or discordant, using a nominal 5% significance threshold to declare discordance (Table 2; Supplemental Data File S4). Concordance was greater for homozygous than segregating

Illumina calls for all variant types, was best for SNPs, and declined with increasing size of insertions and deletions. We conclude that our calls of homozygous SNP and small non-SNP genotypes, which comprise the vast majority of variants, are accurate and that large insertions and deletions should be independently confirmed using other methods.

We compared Freeze 2.0 variants and genotypes with the Freeze 1.0 SNP calls. Of the 5,222,888 polymorphic SNPs in the 158 lines with Freeze 1.0 Illumina data, 4,215,573 are present in the initial Freeze 2.0 call set. The reduction in number of SNP calls was mostly attributable to low frequency SNPs and/or SNPs near indels (Supplemental Fig. S2), suggesting that our integrated variant calling approach eliminated false SNPs near indels. Using a model tailored to the experimental design (Stone 2012), we generated quality scores for each of the 6,149,882 variants and for each genotype in each line. We filtered the genotypes based on the

**Table 2.** Concordance between Illumina and 454 genotyping calls (%)

Type of variant	Size	Homozygous Illumina call		Segregating Illumina call	
		Mean number 454 variants tested/line	% Concordant	Mean number 454 variants tested/line	% Concordant
SNP	N/A	478,049	99.1	59,241	92.7
All non-SNP variants	<100bp	67,467	95.7	36,044	90.6
TR deletion	<100bp	1,077	96.0	1,592	90.9
Non-TR deletion	<100bp	30,465	95.4	13,564	92.8
TR insertion	<100 bp	1055	95.9	1636	86.6
Non-TR insertion	<100bp	30,452	95.9	15,922	90.4
All non-SNP variants	≥100 bp	538	90.4	1354	68.3
CNV deletion	100–400 bp	23	86.1	45	73.8
Non-CNV deletion	100–400 bp	117	94.7	132	88.7
CNV insertion	100–400 bp	24	94.5	45	81.0
Non-CNV insertion	100–400 bp	173	96.7	241	57.5
CNV deletion	>400 bp	57	77.7	291	66.4
Non-CNV deletion	>400 bp	29	78.1	122	65.4
TR insertion	>400 bp	24	76.5	124	76.2
CNV insertion	>400 bp	18	80.5	62	77.8
Non-CNV insertion	>400 bp	56	89.6	90	56.0

quality scores and limited all subsequent analyses to the 4,438,427 biallelic variants meeting the thresholds. For SNPs that were present in both freezes, the concordance rate between the homozygous genotypes was uniformly high (0.9988–0.9996) in all lines.

#### Variation in numbers of segregating sites

The DGRP lines were derived by 20 generations of full-sib inbreeding and have an expected inbreeding coefficient of  $F = 0.986$  (Falconer and Mackay 1996). Therefore, we expect that 1.4% of the variants will remain segregating, under the assumption of selective neutrality. Deleterious variants may be eliminated more rapidly than expected, while an increase in the number of segregating variants could occur from overdominant variants or from de novo mutations. Natural selection favoring heterozygotes can oppose fixation by inbreeding if there is true overdominance for fitness at individual loci or associative overdominance arising from complementary deleterious alleles that are closely linked in repulsion. If complementary deleterious alleles are embedded in polymorphic genetically divergent inversions, inversion heterozygotes may be polymorphic over the entire inverted region. Finally, the appearance of segregating sites can be generated if duplicate, divergent paralogous genes were mapped to a single gene of the pair.

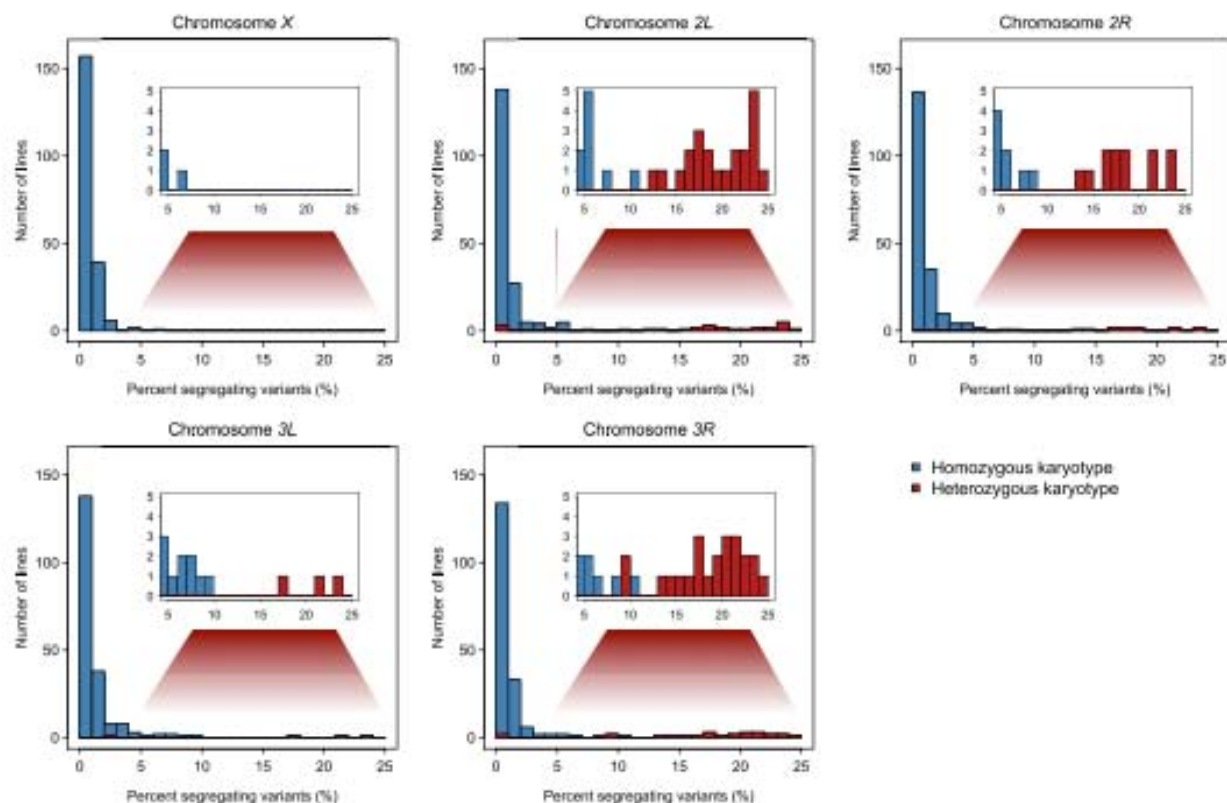
We assessed the number of segregating sites for each line by chromosome (Supplemental Data File S5) and found substantial variation in the number of segregating sites between and within chromosomes. Approximately 96% of the lines had 2% or fewer

segregating X-linked variants, while on average 84% of the lines had 2% or fewer segregating autosomal variants (Fig. 2). Therefore, inbreeding was successful for the majority of variants. However, the distribution of the number of segregating sites on the autosomes was bimodal. In total, 62 of the 820 DGRP line/autosome combinations had  $\geq 9\%$  variants segregating;  $\geq 20\%$  variants remained segregating in 28 chromosomes (Supplemental Data File S5; Fig. 2).

#### Inversion genotypes

*D. melanogaster* populations harbor polymorphic inversions (Stalker 1976; Mettler et al. 1977; Corbett-Detig and Hartl 2012). Recombination is suppressed between the inverted sequence and standard karyotype, leading to divergence between inversions and homo-sequential regions (Navarro et al. 1997, 2000; Andolfatto et al. 2001) and the potential for evolution of coadapted gene complexes (Kirkpatrick and Barton 2006; Hoffmann and Rieseberg 2008). A likely explanation for the large numbers of segregating autosomal variants in specific regions of some lines could be heterozygosity for inversions that are genetically divergent from the standard karyotype. Therefore, we determined inversion genotypes of the DGRP lines by cytogenetic analysis of polytene salivary gland chromosomes.

We identified 16 different segregating autosomal inversions (Table 3; Supplemental Data File S6). Of the 62 autosome/DGRP line combinations with  $>9\%$  segregating sites, 60 had at least one heterozygous inversion, while two were the standard karyotype



**Figure 2.** Distributions of the percent segregating variants in 205 DGRP lines, by chromosome. The distributions for homozygous standard or inverted karyotypes are given in blue, and the distributions for inversion/standard heterozygotes are given in red.

**Table 3.** Inversions in DGRP lines

Inversion	Full name	Chromosome	Cytological breakpoint		Physical breakpoint	
			Start	End	Start	End
<i>In(2L)t</i>	t	2L	22D3-E1	34A8-9	2,225,744 <sup>a</sup>	13,154,180 <sup>a</sup>
<i>In(2R)NS</i>	Nova Scotia	2R	52A2-B1	56F9-13	11,278,659 <sup>a</sup>	16,163,839 <sup>a</sup>
<i>In(2R)Y1</i>	Yutaka#1	2R	49A	55E	8,000,000 <sup>b</sup>	15,000,000 <sup>b</sup>
<i>In(2R)Y2</i>	Yutaka#2	2R	56B	60F	17,000,000 <sup>b</sup>	21,000,000 <sup>b</sup>
<i>In(2R)Y3</i>	Yutaka#3	2R	42A	47E	1,700,000 <sup>b</sup>	7,200,000 <sup>b</sup>
<i>In(2R)Y4</i>	Yutaka#4	2R	51A	56A	10,000,000 <sup>b</sup>	15,000,000 <sup>b</sup>
<i>In(2R)Y5</i>	Yutaka#5	2R	49F	52E	8,900,000 <sup>b</sup>	12,000,000 <sup>b</sup>
<i>In(2R)Y6</i>	Yutaka#6	2R	55E	60F	15,000,000 <sup>b</sup>	21,000,000 <sup>b</sup>
<i>In(2R)Y7</i>	Yutaka#7	2R	53E	56F	12,800,000 <sup>b</sup>	16,200,000 <sup>b</sup>
<i>In(3L)P</i>	Payne	3L	63B8-11	72E1-2	3,173,046 <sup>a</sup>	16,301,941 <sup>a</sup>
<i>In(3L)M</i>	Maurad	3L	66D	71D	8,600,000 <sup>b</sup>	15,000,000 <sup>b</sup>
<i>In(3L)Y</i>	Yutaka	3L	67B	73B	10,000,000 <sup>b</sup>	17,000,000 <sup>b</sup>
<i>In(3R)P</i>	Payne	3R	89C-D	96A	12,257,931 <sup>a</sup>	20,569,732 <sup>a</sup>
<i>In(3R)K</i>	Kodani	3R	86F1-87A1	96F11-97A1	7,576,289 <sup>a</sup>	21,966,092 <sup>a</sup>
<i>In(3R)Mo</i>	Missouri	3R	93D	98F2-3	17,232,639 <sup>a</sup>	24,857,019 <sup>a</sup>
<i>In(3R)C</i>	C	3R	92D1	100F2-3	16,000,000 <sup>b</sup>	26,000,000 <sup>b</sup>

<sup>a</sup>Nucleotide level breakpoints from Corbett-Detig and Hartl (2012).

<sup>b</sup>Approximate physical breakpoints corresponding to cytological map.

(Fig. 2). A possible explanation for the two exceptional karyotypes is that an inversion segregated in these lines when they were sequenced, but the standard karyotype was fixed in the interval between sequencing and the cytological analysis. Of the 758 autosome/DGRP line combinations with fewer than 9% segregating sites, 752 were homozygous for either the inverted or standard sequence (Fig. 2). However, six inversion heterozygotes (one for *In(3L)Y*, two for *In(3R)Mo*, and three for *In(2L)t*) had very low numbers of segregating sites. *In(3L)Y* is only present as a single heterozygote in the sample and could be of recent origin and hence not genetically differentiated from the standard karyotype. However, other chromosomes heterozygous for *In(3R)Mo* and *In(2L)t* had large numbers of segregating sites (Supplemental Data Files S4, S5). Possibly, these inversions do not have a single origin, and the old and new inverted sequences are segregating in the DGRP; or they could have undergone an even number of recombination events as heterokaryotypes, recovering a standard nucleotide configuration. Nevertheless, there is nearly a perfect correlation between large numbers of segregating sites and inversion heterozygosity (Fisher's exact test  $P = 1.91 \times 10^{-83}$ ).

#### Wolbachia infection

*Wolbachia pipientis* is a maternally transmitted endosymbiotic bacterium that infects ~20% of all insects (Dunning Hotopp et al. 2007). *Wolbachia* can manipulate host biology to increase production of infected females, and hence its own transmission (Hoffmann et al. 1986). *D. melanogaster* populations worldwide are polymorphic for *Wolbachia* infection (Richardson et al. 2012). *Wolbachia* infection in *D. melanogaster* has been associated with resistance to infection by RNA viruses (Teixeira et al. 2008), but the full range of effects of *Wolbachia* on development, physiology, reproduction, and quantitative traits is unknown. We determined the *Wolbachia* infection status of the Freeze 2.0 DGRP lines, finding that ~53% of the lines are infected (Supplemental Data File S7). *Wolbachia* sequences have been inserted into eukaryotic genomes (Dunning Hotopp et al. 2007). Therefore, we examined the DGRP lines for evidence of similar lateral gene transfer events and found that all infected lines had predicted insertions of ~180-bp *Wolbachia* sequence at two genomic locations (Supplemental Fig. 3).

However, PCR-based analyses revealed that only four DGRP lines contained the *Wolbachia* insertions (Supplemental Fig. S3). The insertions were incorrectly called in the remaining lines infected with *Wolbachia* because *Wolbachia* sequence reads were present for these lines, and the genotyping algorithm assigned them to the location to which they uniquely mapped in the four lines. This artifact did not occur for any other large insertions, all of which were either unique, as expected for a new *D. melanogaster* sequence present in DGRP lines but not the reference strain, or were homologous to other *D. melanogaster* sequences, as expected from insertions arising from transposable elements (TEs), local tandem duplications, and nonhomologous recombination.

#### Variation in genome size

The large numbers of insertions and deletions suggest that the DGRP lines may vary in genome size. We estimated total genome size for each line using flow cytometry (Hare and Johnston 2011). There is significant variation in genome size (ANOVA  $F_{204, 811} = 2.61$ ,  $P < 0.0001$ ), ranging from 169.7 to 192.8 Mb (Supplemental Fig. S4; Supplemental Data File S8). Genome size differences were verified by the presence of double peaks in copreparations from lines with different average genome size (Ellis et al. 2014). The mean genome size of all lines (175.6 Mb) is close to that of the reference strain (175 Mb). The distribution is skewed toward the accumulation of large genomes, suggesting greater constraint on genome reduction than expansion.

Lines homozygous for *In(2R)NS*, *In(3L)P*, and *In(3R)K* and heterozygous for *In(3L)Y* had larger average genome sizes than the corresponding standard homozygous karyotypes, whereas lines homozygous or heterozygous for all other inversions had smaller average genome sizes than the standard karyotypes. We regressed genome size on the total number of "smaller" inversions and found a significant negative effect ( $b = -0.52$ ,  $F_{1,203} = 8.25$ ,  $P = 0.0045$ ) (Supplemental Fig. S5). Although inversions account for only 4% of the variation in genome size, the magnitude of the effect is substantial at 0.5 Mb per inverted region.

#### Population genomics of indels

Previously, we performed a population genomic analysis of SNPs in the DGRP Freeze 1.0 (Mackay et al. 2012). The SNP genotype calls are highly correlated between Freeze 1.0 and Freeze 2.0. Spearman rank order correlations ( $\rho$ ) for estimates of SNP nucleotide polymorphisms ( $\pi$ ) (Nei 1987) among 100-kb nonoverlapping windows range from  $\rho = 0.94$  for the X chromosome to  $\rho = 0.99$  for 3R (Supplemental Table S1). Since population genomic inferences from analyses of SNP variation remain the same, we primarily focus here on indel variation.

We defined insertions and deletions in our variant calling algorithm with respect to the reference sequence. For population genetic inferences, we polarized insertion/deletion status evolu-



tionarily with respect to *Drosophila simulans* and determined the ancestral and derived status of 210,268 biallelic indels. We found that 86% of "deletions" and 74% of "insertions" inferred from the reference genome were true deletions and insertions according to the polarized estimates.

Evolutionarily derived deletions ( $n = 145,015$ ; 69%) outnumber insertions ( $n = 65,253$ ; 31%) by 2.2:1 (Supplemental Table S2; Supplemental Fig. S6). This estimate is among the highest estimates of the deletion:insertion ratio for *D. melanogaster* but is consistent with previous estimates that indicate a bias toward higher deletion than insertion rates (Petrov 2002; Ometto et al. 2005; Assis and Kondrashov 2012; Leushkin et al. 2013). There are, on average, 60% fewer deletions ( $\chi^2 = 3815$ ,  $P = 0$ ) and 74% fewer insertions ( $\chi^2 = 645.6$ ,  $P = 0$ ) on the X chromosome than on the major autosomal chromosomal arms (Supplemental Table S1), consistent with stronger selection against indels on the X chromosome. The observed bias toward deletions is not an artifact of the greater difficulty of calling large insertions than deletions. We called approximately equal numbers of insertions and deletions except for the largest variants, where we called more deletions than insertions relative to the reference (Table 2). Thus the calling bias is only for variants >400 bp. Since such variants are a very small fraction of the total, this bias cannot account for the excess of evolutionarily derived deletions.

Although most indels are small (1–2 bp), deletions are, on average, larger than insertions (Supplemental Table S2; Supplemental Fig. S6). However, the longest indels are insertions, most of which correspond to *P* transposable elements which have recently colonized the *D. melanogaster* genome (Kidwell 1993). Most large insertions are located in centromeric regions. The distributions of indel size are similar for 3' and 5' UTRs, large and small introns, and intergenic regions, while the size distribution of indels in coding regions has discrete "peaks" for indel sizes in multiples of 3 bp (Supplemental Fig. S7). This pattern suggests strong negative selection against frame-shifting indels compared to more relaxed selection for insertions and deletions spanning complete codons, a phenomenon previously reported for 39 DGRP lines (Massouras et al. 2012) and in humans (Montgomery et al. 2013).

The minor allele frequency (MAF) spectra (Supplemental Fig. S8) show an excess of low MAF indels compared to SNPs for all functional classes. Given that lower MAF variants are likely enriched for variants under purifying selection, these data are consistent with deleterious fitness effects of indels (Massouras et al. 2012). Insertions and deletions causing coding sequence frame-shifts are highly overrepresented among the low derived allele frequency (DAF) class (Supplemental Fig. S9), reinforcing the conclusion that negative selection is intense on this indel class. Relative to presumed neutral variants (synonymous SNPs and SNPs in small introns), all deletion classes have an excess of low-frequency derived alleles on all chromosomes. In contrast, the number of low-frequency derived insertion alleles is similar to or less than presumed neutral SNPs for insertions in small introns and nonframe shifting coding sequence insertions on the X chromosome. There is also a slight excess of high-frequency derived insertions compared to SNPs in all chromosomes and all functional categories except frame-shift insertions. This could indicate more positive selection on insertions than deletions.

These results suggest that natural selection acts differently on insertions and deletions, with stronger purifying selection on deletions (Petrov 2002; Assis and Kondrashov 2012; Leushkin et al. 2013). This is consistent with the mutational equilibrium theory for genome size evolution (Petrov 2002), where optimal genome

size is maintained by purifying selection on small deletions and less selection on long insertions, compensating for sequence loss. This inference from population genomic analysis is consistent with the skewed distribution of genome sizes toward larger genomes.

### Nonrandom distribution of SNPs and indels

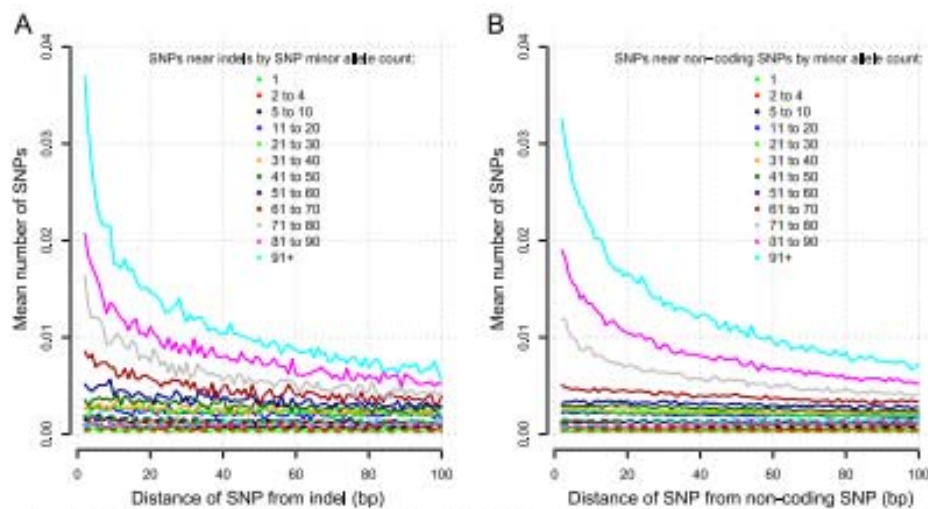
Previously, we found that SNP nucleotide polymorphism ( $\pi$ ) in the DGRP was reduced near centromeres and telomeres and was positively associated with local recombination rate (for recombination rates < 2 cM/Mb) (Mackay et al. 2012). The pattern of  $\pi_{\text{indel}}$  along chromosomes is similar to that of SNP nucleotide diversity (Supplemental Fig. S10). There is a strong positive correlation between indel and nucleotide diversity for all chromosome arms (Supplemental Table S3; Massouras et al. 2012). Several biological mechanisms have been proposed for the clustering of SNPs and indels, which appears to be ubiquitous in prokaryotes and eukaryotes (Tian et al. 2008; Hodgkinson and Eyre-Walker 2011; McDonald et al. 2011; Jovelin and Cutter 2013). Possibly indels (Tian et al. 2008; Jovelin and Cutter 2013) and repeats (McDonald et al. 2011) are mutagenic because they induce error-prone DNA polymerase replication near the indel or repeat (Yang and Woodgate 2007); the regions in which SNPs and indels occur are inherently mutagenic; or SNPs and indels are subject to the same population genomic processes.

To test the hypothesis that indels are mutagenic, we plotted the number of SNPs  $\pm 100$  bp from indels with MAF between 0.4 and 0.5, for different SNP minor allele counts. Intermediate-frequency SNPs are clustered near intermediate-frequency indels (Fig. 3A). Assuming intermediate-frequency indels are older than low-frequency indels, we expect enrichment for SNPs of all minor allele counts near them, since they would continuously generate new mutations. We did not observe this pattern (Fig. 3A). The same analysis for SNPs near intermediate-frequency noncoding focal SNPs also shows an elevated density of SNPs surrounding the focal SNPs (Fig. 3B), indicating that variant clustering is not unique to indel-containing regions. Thus, variant clustering is unlikely to be driven by indels. To test the hypothesis that regions containing increased polymorphism for SNPs and indels have elevated mutation rates, we performed similar analyses for the same regions, but using the lines that do not contain the focal indel alleles. The regions lacking indels contained fewer variants than those with the respective indels (Fig. 3), refuting the locally increased mutation rate hypothesis.

Evolutionary models of hitchhiking and background selection predict a positive correlation between recombination and polymorphism for all variants (Begun and Aquadro 1992; Charlesworth et al. 1993). We replicated our previous observation (Mackay et al. 2012) that SNP polymorphism is positively correlated with the local recombination rate, and extended this observation to insertions and deletions (Supplemental Table S3). Thus, local recombination rate affects the patterning of all types of variants, implicating evolutionary processes as the likely explanation for the observed clustering of variants. The lack of correlation between recombination and divergence for SNPs and indels (Spearman  $\rho = 0.037$  genome-wide,  $P = 0.205$ ) excludes mutations associated with recombination as the cause of the correlation between  $\pi$  and local recombination.

### Distribution of variants in chromatin domains

We determined enrichment or depletion of variants for five chromatin types (Supplemental Data File S9; Filion et al. 2010). Broadly



**Figure 3.** Nonrandom distribution of variants. The average number of SNPs ( $y$ -axis) for each distance in bp ( $x$ -axis) from either side of a variant of high frequency (MAF 40%–50%). Solid lines represent the number of SNPs of a given range of allele counts in lines that have the variant in question, whereas dashed lines show the number of SNPs in lines that do not have the variant. (A) Indels. (B) Noncoding SNPs.

expressed euchromatic genes that perform universal housekeeping functions are depleted of variants, consistent with purifying selection on these genes. Narrowly expressed euchromatic genes associated with more specific biological processes (Filion et al. 2010; van Steensel 2011) are enriched for variants, particularly in coding regions, suggesting that they are under less purifying selection and potentially more rapidly evolving than genes in other chromatin classes. Genes bound by Polycomb Group protein complexes and enriched for the repressive histone mark H3K27me3 are also enriched for variants, which is surprising because Polycomb-associated genes typically regulate developmental processes and are thought to be under strong purifying selection. Genes marked by Heterochromatin Protein 1 binding are located in pericentric regions and are strongly depleted for SNPs and small (<100 bp) indels, but enriched for larger ( $\geq 100$  bp) indels, consistent with our observation that centromeric regions have reduced nucleotide and indel diversity and larger insertions. Interestingly, segmental duplications are highly biased toward centromeric regions in the human genome (She et al. 2004). The most prevalent type of repressive chromatin covers 48% of the genome and marks genes with low expression levels that are generally enriched for variants. While the chromatin classes were derived from one cell type and should be interpreted with caution, our results show that variants are nonrandomly distributed with respect to the chromatin state of the underlying DNA sequence.

#### Population genomics of inversions

Levels and patterning of polymorphism are affected by the recombination landscape and natural selection, both of which are different for regions bearing chromosomal inversions (Navarro et al. 1997; Andolfatto et al. 2001). Recombination is reduced in inversions and is pronounced near the breakpoints of paracentric inversions such that the sequence immediately adjacent to the inversion breakpoint rarely recombines. Recombination is also reduced in inversion heterozygotes because single recombination events within the inverted region lead to inviable aneuploid gametes. However, genetic exchange still occurs in inverted segments from multiple recombination events and/or gene conversion. Thus,

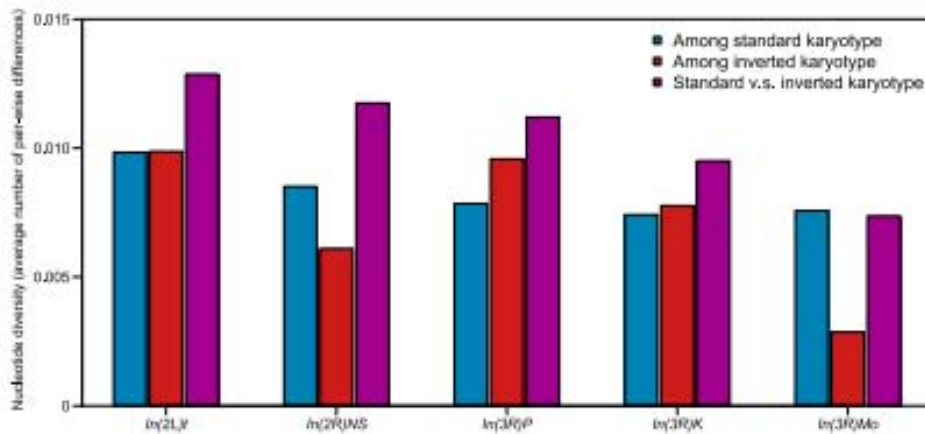
we expect young inversions to have reduced genetic diversity but little divergence from their standard karyotype progenitor, while regions harboring older inversions will separately accumulate mutations in the standard and inverted sequences that lead to differentiation between them. We expect polymorphism to be less within inversion karyotypes, and genetic differentiation to be greater between inversion karyotypes in the regions proximal to the breakpoints than the more central regions of the inverted sequence (Navarro et al. 2000).

Our observation that lines polymorphic for inverted and standard karyotypes have large numbers of segregating sites indeed implies that the inverted and standard karyotypes are genetically divergent. We calculated  $\pi$  for the inverted regions within lines with inverted and standard karyotypes, as well as between the inverted and standard karyotypes (Fig. 4). In all cases, the divergence between karyotypes is higher than the average nucleotide diversity within standard and inverted karyotypes (Fig. 4). However, local variation in polymorphism and diversity swamps any signal of reduction in polymorphism within and increase in diversity between inversion karyotypes near the breakpoints relative to the central regions (Supplemental Fig. S11).

#### Functional annotation of segregating variants

We annotated functional consequences (Supplemental Table S4) of individual segregating variants, identifying 6637 potentially damaging variants that affect splice donor or acceptor sites, cause frame-shift mutations, loss of start or stop codons, or lead to premature stop codons. Collectively, they affect 3868 genes in at least one DGRP line. The allele frequency distribution of these potentially damaging variants is shifted to the lower end of the frequency spectrum relative to those of less damaging variants (Supplemental Fig. S12), as expected if they have deleterious fitness effects.

Next, we identified closely linked cosegregating variants that might ameliorate these potentially damaging variants (Gan et al. 2010). We found pairs of compensatory variants (SNPs that rescue a premature stop codon variant and indels in the same genes that compensate each other to avoid frame-shifts) in an average of



**Figure 4.** Nucleotide diversity ( $\pi$ ) within standard karyotypes (blue bars), within inverted karyotypes (red bars), and between standard and inverted karyotypes (purple bars) within genomic regions encompassed by common polymorphic inversions. The calculation was based on nonmissing genotypes only, with indels (>1 bp) or multiple nucleotide polymorphisms receiving the same weight as SNPs regardless of their length.

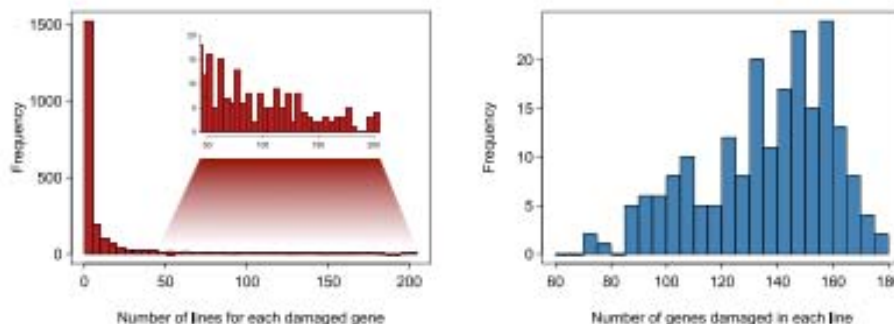
50 genes per line and a total of 403 compensated genes in all lines. These compensatory variants are largely in close physical proximity (1–2 bp) and in near complete linkage disequilibrium ( $D' \sim 1$ ) (Supplemental Fig. S13). In all cases, variants that would otherwise introduce a premature stop codon are present only in lines carrying the compensatory variants. Given their close proximity, recombination events are unlikely to occur between pairs of adjacent compensated variants. This suggests that the compensatory variants at these codons most likely occurred first in the population, thus allowing the second mutation to occur without introducing a stop codon. Consistent with our inferred timeline of mutations, these compensated variants segregate at higher frequency in the DGRP than other potentially damaging variants (Supplemental Fig. S14).

Finally, we performed gene-centric annotation by integrating all sequence variations overlapping coding regions in each DGRP line to take into account the widespread occurrence of multiple variants in single genes. We found 2169 genes whose proteins are damaged by the combination of all variants in them in at least one DGRP line (~15% of *Drosophila* protein coding genes) (Supplemental Data File S10). On average, each of these affected genes is damaged in ~13 of the 205 DGRP lines, and each line contains ~136 potentially damaged genes (Fig. 5). These potentially damaging variants and genes are a new source of novel mutations for functional analyses. Gene ontology enrichment analysis showed that

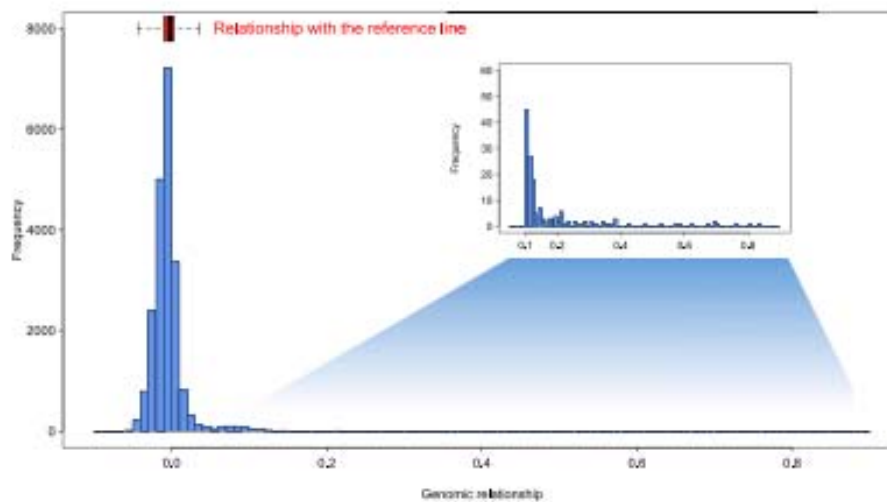
multigene families affecting chemosensation, detoxification of xenobiotic substances, immune and defense response, and proteolysis are enriched for damaged genes (Supplemental Data File S11). The same gene families are rapidly evolving along the *Drosophila* phylogeny (*Drosophila* 12 Genomes Consortium 2007).

#### Genetic relationships among DGRP lines

Genetic diversity is highly elevated between inverted and standard karyotypes in the region of the inversion. Thus, we expect that individuals of the same inversion karyotype will be more related to each other than to individuals of the standard karyotype. Therefore, we quantified patterns of genetic relatedness among the DGRP lines by constructing the genetic relationship matrix between all pairs of DGRP lines (Supplemental Fig. S15; Van Raden 2008; Ober et al. 2012). The distribution of relatedness is bimodal with the major peak centering around zero and the vast majority of pairwise relatedness within the range of distance to the reference strain (Fig. 6). The minor peak consists of 567 pairs (2.7% of all possible pairs) with relatedness greater than 0.05. There are 11 pairs (0.05% of all possible pairs) among 16 lines that have a genomic relationship greater than 0.5. Therefore, most DGRP lines are unrelated, consistent with sampling from a large, randomly mating population. However, some lines have higher genomic relatedness due to cryptic genetic relatedness (Astle and Balding



**Figure 5.** Histograms of the numbers of DGRP lines containing each damaged gene (left) and the number of damaged genes per DGRP line (right).



**Figure 6.** Histogram of genomic relationships among DGRP lines (20,910 possible pairs). The distribution of the relationship between all DGRP lines and the reference sequence is displayed as a box plot.

2009), possibly caused by sampling siblings from the natural population and/or shared inversion karyotypes.

Principal component (PC) analysis reveals clusters of related lines that carry major inversions. The first two PCs separate lines carrying both *In(2L)t* and *In(3R)Mo* from all other lines (Fig. 7A), while the first and third PCs discriminate lines with *In(2L)t* from those with *In(3R)Mo* (Fig. 7B). The PC clustering by inversions disappears when variants within the inverted regions are excluded (Fig. 7C). Lines with the same inversions are more related to each other than are lines homozygous for the standard karyotype (Supplemental Fig. S16), confirming that the PC clusters are driven by increased average genomic relationships within inversions.

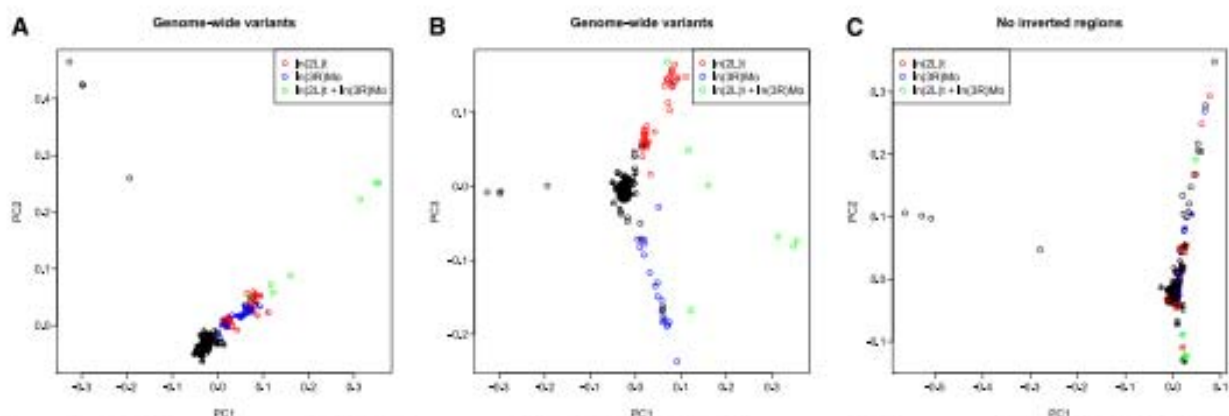
We also computed genomic relationships separately for each chromosome arm (Supplemental Fig. S17). The chromosome-wide relationships among the lines are specific to each arm and are different from the genome-wide pattern (Supplemental Fig. S17). The genomic heterogeneity of relatedness among chromosomal arms suggests that population structure other than the known

inversions is likely minimal; otherwise, inter-chromosomal correlation of relatedness would arise.

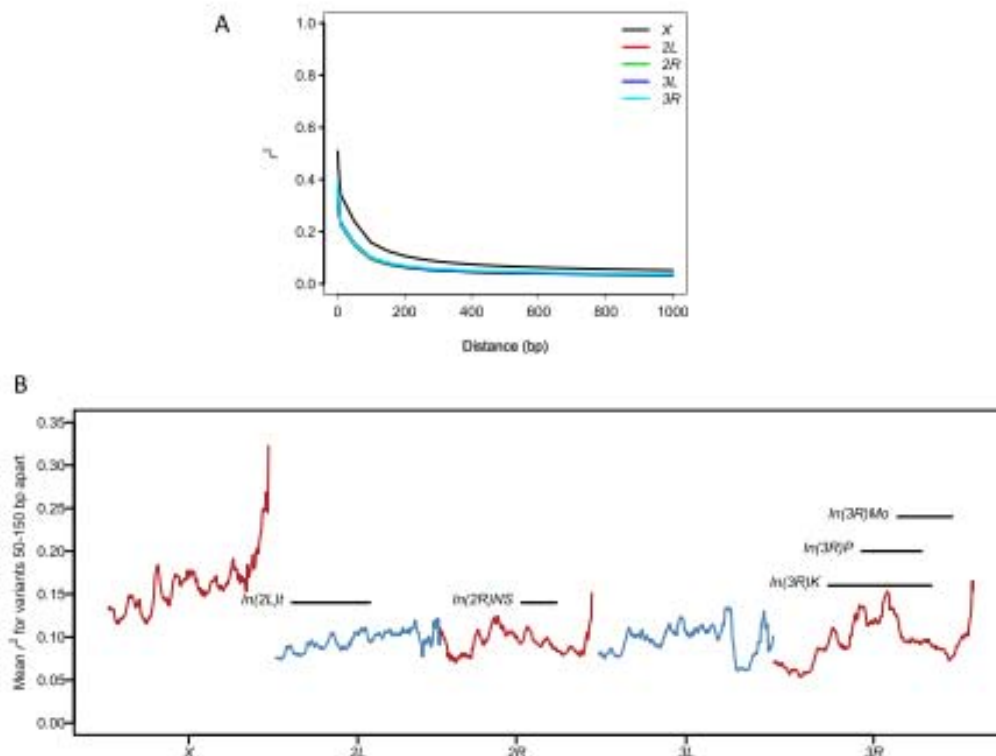
#### Linkage disequilibrium

We assessed pairwise linkage disequilibrium (LD) between polymorphic variants using the  $r^2$  parameterization (Hill and Robertson 1966). Average LD decays rapidly as the distance between the variants increases, and the rate of decay is substantially lower on the X chromosome than autosomes (Fig. 8A), consistent with previous observations based on fewer DGRP lines and SNP variants only (Mackay et al. 2012). There is substantial variation in local LD along the genome (Fig. 8B). In general, LD near centromeres and telomeres is significantly greater than in other chromosomal regions.

The rapid decline in local LD with physical distance is favorable for identifying causal genes and possibly variants in genome-wide association (GWA) studies using the DGRP. However, long-range LD could significantly impair our ability to identify



**Figure 7.** Principal component analysis of DNA sequence variation in the DGRP. Principal components (PCs) are computed using EIGENSTRAT. (A) PC plot of PC1 versus PC2. (B) PC plot of PC1 versus PC3. (C) PC plot of PC1 versus PC2 after PCs were recomputed excluding all variants in regions encompassing major inversions (*In(2L)t*, *In(2R)NS*, *In(3R)P*, *In(3R)K*, *In(3R)Mo*). With the exception of four highly related pairs of lines, there is no apparent clustering of karyotype groups.



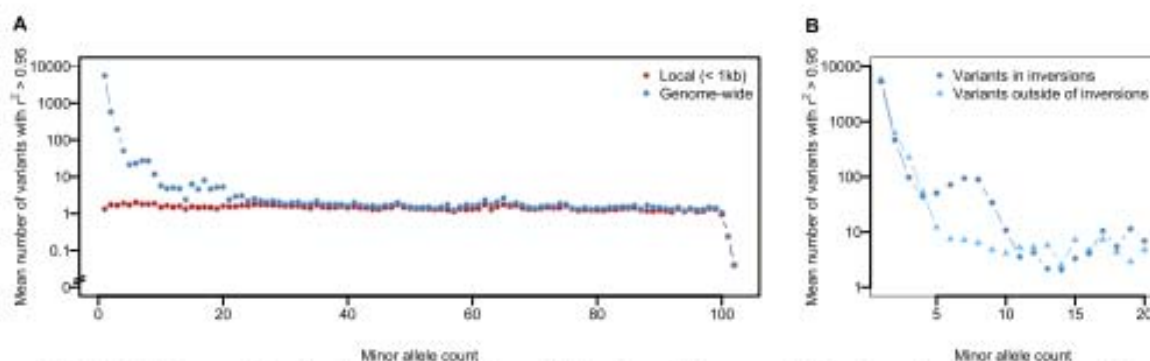
**Figure 8.** Patterns of LD. (A) Decay in LD with physical distance, by chromosome arm. (B) Genome-wide spatial variation in LD. Mean  $r^2$  between variants within 50–150 bp of each other in sliding windows (in 100-kb steps) of 1 Mb is plotted.

QTLs. For each of 1000 randomly sampled variants with a specified number of minor alleles in the population, we counted variants that are in strong LD ( $r^2 > 0.95$ ) with it locally (within 1 kb) and genome-wide. There are consistently very few (mean = 1.43) variants in high local LD with the focal variant. However, the number of long-range variants in high LD with focal variants depends on the minor allele count of the focal variant and can be in the thousands for very low frequency variants (Fig. 9). Although local LD does not seem to differ for the regions with or without inversions (Fig. 8B), long-range LD as measured by the number

of nonlocal variants in high LD is greater for variants within inversions (Fig. 9). Therefore, GWA studies based on individual variants should be restricted to common polymorphisms and also take into account inversions.

#### Associations between quantitative traits, *Wolbachia*, inversions, and genome size

The range and magnitude of effects of *Wolbachia* infection and segregating inversions on organismal phenotypes is not known.



**Figure 9.** Relationship between LD and minor allele count. For each of the minor allele counts, 1000 random variants are sampled, and the mean number of variants genome-wide or locally (<1 kb) in strong LD ( $r^2 > 0.95$ ) with the focal variant is calculated. (A) Relationship between the mean number of variants in strong LD with the focal variant and minor allele count. (B) Relationship between the mean number of variants in strong LD with the focal variant and minor allele count, stratified according to the location of the focal variant (within or outside of inversions).

Therefore, we assessed to what extent inversion genotypes and *Wolbachia* infection status are associated with starvation resistance, startle response, time to recover from chill coma (Mackay et al. 2012), resistance to acute (Weber et al. 2012) and chronic (Jordan et al. 2012) oxidative stress, several sleep phenotypes (Harbison et al. 2013), and olfactory behavior (Swarup et al. 2013). The effect of *Wolbachia* is only significant for acute and chronic resistance to oxidative stress (Supplemental Table S5). *In(3R)K* is associated with starvation resistance in females and acute oxidative stress resistance in males; *In(2L)t*, *In(2R)NS*, and *In(3R)Mo* are associated, often strongly and in a sex-specific manner, with sleep traits; and *In(2L)t* and *In(3R)Mo* are associated with olfactory behavior in both sexes (Supplemental Table S5). The DGRP lines vary significantly in genome size, which could also affect variation in quantitative trait phenotypes. However, correlations of quantitative traits with genome size were small for all traits and not significant in any analysis.

### Genome-wide association analyses in the DGRP

Prior to performing GWA analyses using the DGRP, we must adjust the phenotypic data to account for cryptic genetic relatedness, effects of inversions (lines with the same inversion karyotype have higher relatedness, and there is elevated LD within inverted regions) and *Wolbachia* infection status. Association tests can be performed for individual variants or by gene. The former can identify putative causal alleles but is restricted to the 1,920,276 variants with minor allele frequencies  $\geq 0.05$  to avoid spurious associations due to LD caused by limited sample size (Fig. 9). Gene-based tests can interrogate the remaining variants with low allele frequencies, which should contribute substantial variation if variation in the trait is maintained by mutation-selection balance (Turelli 1984), and can also evaluate effects of common variants and all variants. However, they are sensitive to the exact methods used for weighting variants within a gene (Madsen and Browning 2009; Han and Pan 2010; Wu et al. 2010, 2011; Lee et al. 2012). In either scenario, we perform associations on the adjusted phenotypic values using a model that accounts for cryptic relatedness among the lines. For single marker association, we use a mixed model that incorporates the relationship matrix, whereas for the gene-based tests, we add covariates corresponding to the major principal components that account for relatedness. We performed GWA analyses for starvation resistance, a classic quantitative trait, *Wolbachia* infection status (in this case, the data were not corrected for *Wolbachia* infection), and genome size (Supplemental Data Files S12, S13; Supplemental Text S1).

The need to adjust for *Wolbachia* and inversions and account for relatedness is illustrated by quantile-quantile plots (Supplemental Fig. S18) from single variant GWA analysis of starvation resistance in females, which is associated with *In(3R)K* (Supplemental Table S5). Unadjusted data show substantial systematic inflation of test statistics, while adjusting for *Wolbachia* and inversions and accounting for relatedness using a mixed model significantly alleviate the inflation. The top associations for the individual and gene-based tests for all three traits are only partially overlapping, highlighting the complementary nature of these tests. Only a few variants/genes reached conservative Bonferroni-adjusted significant thresholds, and all suggest novel candidate genes affecting the traits. Examples include a SNP in *genghis khan* (*gek*, a protein kinase), associated with female starvation resistance, and SNPs in *pointed* (*pnt*, a transcription factor) and *CG32521* (a gene of unknown function), associated with genome size.

*myotubularin* (*mtm*), which is involved in chromosome segregation and the mitotic cell cycle (McQuilton et al. 2012), is a plausible candidate gene associated with genome size and reached Bonferroni-level significance in the gene-based tests of association with this trait.

### Discussion

Here, we present a molecular polymorphism map for 205 sequenced inbred *D. melanogaster* lines comprising Freeze 2.0 of the DGRP. We utilized seven different algorithms for detecting variants to produce a consensus variant list, and further fine-tuned the variant calls for each line using an integrated genotyping strategy that borrows power from the variant calls in all lines. We further provide quality scores for all 4,853,802 SNP and 1,296,080 non-SNP variants using a method that takes into account the experimental design used to generate the DGRP. Independent validation of variant calls gives low false positive rates for SNPs and small (<100 bp) indels, which comprise >98% of all variants. We performed a cytogenetic analysis of large segregating inversions, genotyped all lines for the presence of the maternally transmitted *Wolbachia* endosymbiont, and estimated genome size by flow cytometry. These data provide a comprehensive characterization of natural variation in genome architecture in this powerful genetic model organism that can be used to gain insights about natural selection and the evolution of genome size, and enhance the functional annotation of the *D. melanogaster* genome. We also describe improved statistical methodology for genome-wide association mapping of quantitative traits in a scenario where all variants are known and the rapid decay in LD with physical distance enables high-resolution mapping.

Our molecular population genomic analysis of evolutionarily polarized deletion and insertion variants showed that deletions outnumber insertions by a ratio of greater than 2:1, consistent with previous studies on smaller data sets, suggesting a bias toward the deletion mutation rate in *Drosophila* (Petrov 2002; Assis and Kondrashov 2012; Leushkin et al. 2013). Site frequency spectra show an excess of low-frequency polymorphisms compared to SNPs for insertions and deletions from all functional categories but especially for frame-shifting indels, implicating strong purifying selection against these variants. However, the site frequency spectra suggest stronger selection on deletions than insertions, which could lead to the maintenance of an optimal genome size (Petrov 2002). Our direct observation of variation in genome size in the DGRP, which varies by ~14%, is in accord with this hypothesis. This variation in genome size is similar to that observed for an *Arabidopsis thaliana* population in Sweden (Long et al. 2013). The distribution of genome size variation is skewed toward larger genomes, consistent with stronger purifying selection against deletions than insertions.

As observed previously (Tian et al. 2008; McDonald et al. 2011; Massouras et al. 2012; Jovelin and Cutter 2013), we found a strong positive correlation between the genomic distribution of indels and SNPs. These correlated patterns of polymorphism are, in turn, correlated with local recombination, suggesting that the nonrandom distributions are due to hitchhiking and background selection (Begun and Aquadro 1992; Charlesworth et al. 1993). Alternative explanations that indels are mutagenic or that the highly polymorphic regions have high mutation rates were not supported by our analyses.

Inversions are islands of genomic divergence in this *D. melanogaster* population. Nucleotide diversity is elevated between

inverted and homo-sequential genomic regions relative to the average diversity of inverted and standard regions, and consequently, lines heterozygous for inversions have large numbers of segregating sites in the region encompassed by the inversion. There is a greater extent of long-range LD within inverted sequences than the same regions on the standard karyotypes, indicative of lower recombination rates and effective population sizes of inversions. It is intriguing that variation in genome size is significantly associated with inversions. The mechanistic basis of increased or decreased genome size in the different inversion karyotypes is an open question for future study. Previously, we inferred that there was little global population structure in the DGRP from our eigen-decomposition of the genetic covariance matrix, but noted that the large variance in this decline did not preclude local structure due to structural variation (Mackay et al. 2012). Here, we performed a more comprehensive analysis of variation in genetic relatedness in the DGRP and showed that individuals with the same inversion karyotype are more related to each other than to individuals of the standard karyotype, accounting for most of the variation in relatedness among the DGRP lines and local structure. Inversions can harbor “coadapted gene complexes” associated with fitness (Dobzhansky 1937), and indeed, many fitness-related traits have been associated with inversion polymorphism in *Drosophila* species (Hoffmann and Rieseberg 2008). We showed that variation in starvation and oxidative stress resistance, sleep traits, and olfactory behavior are all associated with inversion polymorphism, and future evaluation of more quantitative traits in the DGRP will provide a detailed picture of effects of inversions on complex traits.

Lateral gene transfer of *Wolbachia* sequences into insect genomes is common, most likely because its presence in developing gametes is a favorable scenario for germline integration (Dunning Hotopp et al. 2007). Lateral gene transfer is a potential mechanism for the acquisition of novel genes, but to date has not been reported for *Wolbachia* sequences in *D. melanogaster*. We identified two different insertions of small *Wolbachia* insertions in four DGRP lines. Future analyses of the transcriptomes of these lines will reveal whether the insertions are transcribed and potentially functional. The forces maintaining *Wolbachia* infection in *D. melanogaster* populations near 50% remain mysterious. Although infection status has been associated with resistance to infection by RNA viruses (Teixeira et al. 2008), effects of *Wolbachia* infection on the quantitative traits assessed in the DGRP are rarely significant.

The goal of the Berkeley *Drosophila* Genome Project (BDGP) Gene Disruption Project (Bellen et al. 2004, 2011) is to generate mutations in all *D. melanogaster* genes as tools for functional analysis, and that of the *Drosophila* modENCODE Project (The modENCODE Consortium et al. 2010) is to identify sequence-based functional elements in *Drosophila*. The DGRP complements these efforts. The millions of molecular variants segregating in the DGRP are novel mutations for functional analysis and represent a different functional class from the transposon-tagged mutations produced from the BDGP Gene Disruption Project. Indeed, 15% of all *D. melanogaster* genes segregate for potentially damaged proteins in the DGRP, yet these damaged genes are compatible with live, fertile flies (at least under standard laboratory conditions). Molecular population genomic analyses using the DGRP highlight genomic regions under purifying selection, complementing modENCODE functional motifs. GWA analyses of quantitative traits in the DGRP provide new functional annotation of the *D. melanogaster* genome by identifying novel candidate genes asso-

ciated with these traits. These genes typically have well-described effects on other traits, play key roles in early developmental processes, or are computationally defined genes with no known function, but have never been associated with the focal trait. Subsequently, the full power of *Drosophila* genetics can be applied to validating marker-trait associations: mutations, RNAi constructs, and outbred QTL mapping populations (Huang et al. 2012; Jordan et al. 2012; Weber et al. 2012; Harbison et al. 2013; Swarup et al. 2013). The future of understanding the genetic architecture of quantitative traits lies in our ability to progress from one-gene-at-a-time associations to understanding how entire genetic and transcriptional networks causally affect complex organismal phenotypes. The DGRP is an ideal resource for systems genetics (Ayroles et al. 2009; Massouras et al. 2012) and epistatic interaction network analyses (Yamamoto et al. 2009; Huang et al. 2012; Swarup et al. 2013) of molecular and complex organismal traits.

The DGRP lines, sequence data, genotypes, quality scores, and phenotypes are publicly available. The DGRP website (<http://dgrp2.gnets.ncsu.edu>) hosts an updated pipeline for single marker GWA analysis which accounts for effects of *Wolbachia* infection and major inversions as well as cryptic relatedness among the DGRP lines; a new genome browser track for visualizing individual line genotypes and functional annotations for any specified genomic region; and all published phenotypes. These data will be useful for testing new analytical methods as well as for teaching general principles of population and quantitative genetics.

## Methods

### DGRP lines

We established isofemale lines from gravid females collected in Raleigh, NC, and inbred them by 20 generations of full-sib mating, followed by random mating (Mackay et al. 2012). All flies were reared and all phenotypes assessed under standard culture conditions (cornmeal-molasses-agar-medium, 25°C, 60%–75% relative humidity, 12-h light-dark cycle) unless otherwise specified.

### DNA isolation, library construction, and sequencing

We extracted genomic DNA from ~500–1000 flies per DGRP line using the Genra Puregene Tissue Kit (Qiagen) and purified the samples by phenol-chloroform extraction. We constructed high molecular weight double-strand genomic DNA samples into Illumina paired-end libraries according to the manufacturer's protocol (Illumina) (Supplemental Text S2) and sequenced shotgun DNA libraries on the Illumina HiSeq 2000 or GAII platforms, according to the manufacturer's specifications.

### Sequence read mapping and initial genotyping

We aligned Illumina sequence reads to the Dmel 5.13 reference genome (<http://flybase.org>) with BWA (v0.5.9-r16) (Li and Durbin 2010) and Novoalign (Novocraft.com) using default parameters. We used GATK (v1.0.5506) (McKenna et al. 2010) software to remove duplicate sequence reads, recalibrate base quality scores, and locally realign regions around indels for BWA alignments (DePristo et al. 2011). We excluded positions with >2000 coverage and mapped reads with *phred* scores <25 and/or mapping quality <10. We applied GATK (v1.0.5506) (McKenna et al. 2010) and JGIL (Stone 2012) to the BWA and Novoalign alignments, and Atlas-SNP (Shen et al. 2010) and PrinSeS (Massouras et al. 2010) to the BWA alignments to genotype SNPs. We genotyped

non-SNP variants <100 bp using GATK, Atlas-SNP, and PrinSeS. We genotyped non-SNP variants  $\geq 100$  bp using PrinSeS, DELLY (Rausch et al. 2012), Pindel (v0.2.4d) (Ye et al. 2009), CNVnator (v0.2.2) (Abyzov et al. 2011), and Genome STRIP (v1.0.4) (Handsaker et al. 2011) as described in Zichner et al. (2013).

### Integrated genotyping

We performed integrative genotyping in two stages. First, we genotyped each line separately using all SNP and non-SNP variants from the output of the individual variant calling methods to provide the alternative haplotypes from which to choose variants. In the second stage, we again performed genotyping for each line, using the 205 variant lists resulting from the first stage (Supplemental Text S2). The resulting 6,149,882 nonredundant variants were then assigned variant and genotype quality scores using JGIL (Stone 2012). We retained for subsequent analyses nonoverlapping biallelic variants whose *phred* scale quality scores were at least 500 and genotypes whose sequencing depths were at least one and genotype quality scores at least 20. The final VCF genotype file (<http://dgrp2.gnets.ncsu.edu>) containing 4,438,427 variants gives the number of supporting and opposing reads for each variant in each line, genotypes with the maximum posterior probability, and the corresponding quality scores (Stone 2012).

### Validation

We used three strategies to validate genotype calls. First, we performed Sanger sequencing for 384 small (1–18 bp) indels affecting coding regions and 384 larger (30–313 bp) randomly chosen indels on five DGRP lines (DGRP\_304, DGRP\_324, DGRP\_354, DGRP\_355, DGRP\_395). Second, we used previously published data (Zichner et al. 2013) on genomic DNA hybridization to Affymetrix GeneChip *Drosophila* 1.0R tiling arrays for six DGRP lines (DGRP\_208, DGRP\_304, DGRP\_313, DGRP\_315, DGRP\_437, DGRP\_555) and the reference strain to validate deletions >25 bp (Supplemental Text S2). Finally, we used 454 sequence (Roche) data from 38 DGRP lines (Mackay et al. 2012) to validate SNP and non-SNP calls (Supplemental Text S2). We used our integrated genotyping algorithm to count supporting and opposing reads of alleles for variants and tested the allele counts from Illumina and 454 for concordance using a Fisher's exact test.

### Inversion karyotypes

We assessed inversion genotypes by cytogenetic analysis of polytene salivary gland chromosomes of third instar larvae by staining with lactic-acetic orcein. We identified inversions by comparison to the standard map of Bridges (1935). We initially examined two larvae from each DGRP line and subsequently confirmed inversion heterozygotes or segregating inversions by examining additional larvae and/or F1 hybrids of the DGRP line with the standard Canton S karyotype.

### Wolbachia status

We used PCR to determine the infection status of each line with respect to the endosymbiont, *Wolbachia pipientis* (Braig et al. 1998; Richardson et al. 2012; Supplemental Text S2). We used DGRP\_101 and DGRP\_105 as negative controls and DGRP\_142 and DGRP\_149 as positive controls. We also developed PCR assays to genotype all DGRP lines for insertions of *Wolbachia* genome at 2R:16,594,660 and 2R:19,117,791 (Supplemental Text S2). We purified PCR products for lines positive for *Wolbachia* insertions using the Zymo Clean and Concentrator kit (Zymo Research Corporation)

and subjected them to Sanger sequencing using the ABI 3730XL platform.

### Genome size

We estimated genome sizes for 1016 individual females (at least three individuals per line) using flow cytometry with *Drosophila virilis* (1C = 328 Mb) as an internal standard, as described in Hare and Johnston (2011) but with a final concentration of propidium iodide stain at 25 mg/mL. The estimate of genome size was the proportion of stain uptake (expressed as a channel number by the flow cytometer) of the sample relative to the standard times the amount of DNA in the standard. We calculated the average genome size and standard deviation of genome size and performed additional replicate measurements as needed to produce a standard error of 0.5%. We tested whether the differences in genome sizes were true by flow cytometry analysis of copreparations of females from lines with different average genome size. We evaluated the association of segregating inversions with variation in genome size using the ANOVA model  $Y = \mu + G + \epsilon$ , where  $Y$  is the standard deviation of genome size within a line and  $G$  is the number of segregating inversions (0, 1, 2, 3, or 4) within lines.

### Population genomics

We used 357,708 JGIL-filtered, biallelic indels present in at least 101 lines to conduct the indel population genomics analyses. We assigned indels to one of six functional classes (coding sequence, 5' and 3' UTR, long [ $>100$  bp] and short [ $\leq 100$  bp] introns, intergenic sequence) using the 5.49 version annotations of the *D. melanogaster* reference genome (Marygold et al. 2013). We discarded indels spanning more than one functional class, leaving 357,608 indels with a valid functional class. We analyzed insertions and deletions separately, after first polarizing ancestral and derived states with respect to the high quality second-generation assembly genome of *D. simulans* (Hu et al. 2013) as an outgroup (Supplemental Text S2). We inferred the derived allele status for 210,268 indels. We manually checked a sample of 500 derived indels to which our polarizing protocol was applied; all were correct. Therefore, we conclude that the specificity of our procedure is very high, although we excluded 41% of the original indel data set from our evolutionary analyses.

We used  $\pi_{indel}$  to describe indel polymorphism, a measure analogous to nucleotide diversity ( $\pi$ ), which does not take into account indel size. We used an analogous measure to estimate divergence ( $k$ ) (Librado and Rozas 2009). We estimated fixed indel divergence for *D. melanogaster*, *D. simulans*, and *D. yakuba* using the multiple alignments *D. melanogaster* Oct. 2006 from the VISTA Browser (Frazer et al. 2004). We estimated these diversity measures for the whole genome and by chromosome arm (*X*, *2L*, *2R*, *3L*, *3R*, *4*) in 100-kb nonoverlapping windows. We also estimated the minor allele frequency (MAF) distribution for indels and the derived allele frequency (DAF) distributions for both deletions and insertions. We used the nonparametric Spearman rank correlation coefficient ( $\rho$ ) to test for covariation among the diversity estimates. We used the recent high resolution recombination map of *D. melanogaster* (Comeron et al. 2012) to correlate recombination with the diversity measures.

### Functional annotation of variants

We annotated the functional consequences of variants on annotated genes (FlyBase R5.49) (Marygold et al. 2013) using SnpEff (v3.1m) (Cingolani et al. 2012). We considered variants annotated



as SPLICE\_SITE\_ACCEPTOR, SPLICE\_SITE\_DONOR, START\_LOST, FRAME\_SHIFT, STOP\_GAINED, STOP\_LOST to be “potentially damaging” for the affected proteins. We also performed a line-specific annotation integrating all homozygous variants each line carries. For each gene, we translated the variant transcript using the standard genetic code and compared the variant protein to the reference protein using the global alignment “stretcher” utility in EMBOSS (v6.5.7) (Rice et al. 2000). We considered the variant protein to be potentially damaged if the START or STOP codon was lost or the sequence identity with the reference protein was smaller than 90%. We considered a gene to be potentially damaged if all of its splice variants were affected.

#### Analysis of relatedness and population structure

We calculated the realized genome-wide relationship matrix  $\mathbf{G}$  among all DGRP lines using biallelic common variants ( $\text{MAF} > 0.05$ ) with a call rate  $>80\%$ . This computation was performed using the Van Raden (2008) formula implemented in the rrBLUP R package (v4.0) (Endelman 2011). The relationship matrix was normalized by the mean value of the diagonal elements. For analysis of population structure, we performed a principal component analysis (PCA) using EIGENSTRAT (v4.2) (Price et al. 2006). We pruned LD using the LD pruning utility in PLINK (v1.07) (Purcell et al. 2007) such that in a moving window of every 500 variants, the maximum pairwise  $r^2$  was smaller than 0.2. We excluded variants within 2 Mb of the major inversions (2L:0.4Mb-14.9Mb, 2R:9Mb-18Mb, 3R:6Mb-27Mb) from this analysis. We tested the significance of the top eigenvalues using the Tracy-Widom statistic implemented in EIGENSTRAT.

#### Variant-based association mapping

We performed genome-wide association studies in two stages. In the first stage, we adjusted the data for the effects of *Wolbachia* infection and major inversions [*In(2L)t*, *In(2R)NS*, *In(3R)P*, *In(3R)K*, and *In(3R)Mo*] based on mean phenotypic values of each line. We then used the adjusted line means to fit a linear mixed model in the form of  $y = \mathbf{X}b + \mathbf{Z}u + e$ , where  $y$  is the adjusted phenotypic values,  $\mathbf{X}$  is the design matrix for the fixed SNP effect  $b$ ,  $\mathbf{Z}$  is the incidence matrix for the random polygenic effect  $u$ , and  $e$  is the residual. The vector of polygenic effects  $u$  has a covariance matrix in the form of  $\mathbf{A}\sigma^2$ , where  $\sigma^2$  is the polygenic variance component. We fitted this linear mixed model using the FastLMM program (v1.09) (Lippert et al. 2011).

#### Gene-based association mapping

We performed a burden test and a nonburden sequence kernel association test (SKAT) to assess the cumulative effect of all variants within one kilobase of each annotated gene. The weighted burden test weights the contribution of each variant in a gene by the reciprocal of the standard deviation of its estimated minor allele frequency and uses the weighted averages to estimate a score statistic (Madsen and Browning 2009; Han and Pan 2010). The SKAT kernel function builds a relationship matrix detailing relatedness of individuals based upon all variants within a gene. This relationship matrix is fit as the covariance matrix of a random effect in a linear mixed model framework and used to estimate a variance component score to discern the significance of a trait association (Wu et al. 2011). The SKAT kernel function used was linear and did not up-weight the relative contribution of minor alleles.

We performed both the weighted burden test and SKAT using the SKAT package (Wu et al. 2011) in R v3.0.1 (R Development

Core Team 2013). For both methods, male and female starvation resistance and genome size were fit with an identity link function and fixed effect covariates for *Wolbachia* infection status, major inversions, and the 11 principal components explaining the most genetic variation in the DGRP (Tracy-Widom  $P$ -value  $< 0.01$ ). *Wolbachia* infection status was fit with a logit link function in a likewise manner, excluding the fixed effect of *Wolbachia* infection status. We performed gene-based tests for all variants, and for common ( $\text{MAF} \geq 0.05$ ) and rare ( $\text{MAF} < 0.05$ ) variants separately.

#### Data access

The DGRP lines are available from the Bloomington *Drosophila* Stock Center (<http://flystocks.bio.indiana.edu/Browse/DGRP.php>) (see Supplemental Data File S1 for stock numbers). Raw sequence data have been submitted to the NCBI Sequence Read Archive (SRA; <http://www.ncbi.nlm.nih.gov/sra>) under accession numbers listed in Supplemental Data File S1, and to the Baylor College of Medicine Human Genome Sequencing Center (<https://www.hgsc.bcm.edu/arthropods/drosophila-genetic-reference-panel>). The genotypes, quality scores, phenotypes, and web-based analysis tools are available from the DGRP website (<http://dgrp2.gnets.ncsu.edu>).

#### List of affiliations

<sup>1</sup>Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina 27595, USA; <sup>2</sup>Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland; <sup>3</sup>Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; <sup>4</sup>Center for Education in Liberal Arts and Sciences, Osaka University, Osaka-fu, 560-0043 Japan; <sup>5</sup>Genomics, Bioinformatics and Evolution Group, Institut de Biocnologia i de Biomedicina (IBB), Department of Genetics and Microbiology, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain; <sup>6</sup>Department of Entomology, Texas A&M University, College Station, Texas 77843, USA; <sup>7</sup>Genome Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany; <sup>8</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030 USA; <sup>9</sup>Virginia Tech Virginia Bioinformatics Institute and Department of Biological Sciences, Virginia Tech, Virginia 24061, USA.

#### Acknowledgments

The work was supported by NIH grants NHGRI U54 HG003273 (R.A.G.), GM R01 GM45146 (T.E.C.M., R.R.H.A., E.A.S.), R01 GM076083 (T.E.C.M., R.R.H.A., E.A.S.), R01 AA016560 (T.E.C.M., R.R.H.A.), and R01 GM 59469 (R.R.H.A., T.E.C.M.); the Swiss National Science Foundation grant CRSI33\_127485 (B.D.); institutional support from the École Polytechnique Fédérale de Lausanne (EPFL) and VITAL-IT for computational analyses (B.D.); the German Research Foundation Emmy Noether Fellowship grant KO 4037/1-1 (J.O.K., T.Z.); NSF grant REU-BIO-1062178 (C.E.H.); National Institute of Justice Grant 2012-DN-BX-K024 (A.M.T.); Spanish Ministerio Ciencia e Innovación grant BFU2009-09504 (A.B.); and startup funds from Texas AgriLife Research and the Texas A&M University College of Agriculture and Life Sciences (A.M.T.). Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice.

*Author contributions:* Conceived the project: T.E.C.M., R.A.G., S.R.; Illumina sequencing: S.L., J.J., K.B., C.W., Y.M.Z., Y.Q.Z., M.M., L.L.P., L.P., N.S., S.P., Y.Q.W., Y.H., M.J., F.O., D.K., M.A.C.;

variant discovery: A.M., T.Z., D.Z., G.H., J.O.K., D.M., E.A.S., B.D., T.F.C.M.; genotyping: A.M., W.H., E.A.S., B.D., T.F.C.M.; Sanger validation: S.F., L.L., A.P., A.W., R.R., K.C., S.R.; population genomics: A.B., M.R., W.H., A.M., B.D.; inversion karyotypes: Y.L., A.Y.; *Wolbachia* analysis: L.T., M.M.M., W.H.; genome size analysis: J.S.J., A.M.T., L.L.E., C.E.H.; functional annotation: W.H.; population structure: W.H., E.A.S.; starvation resistance data: S.M.R.; GWA analyses: W.H., J.P., M.M.M.; DGRP construction, maintenance, and Bloomington *Drosophila* Stock Center liaison: R.E.L.; website development and implementation: W.H., J.R.J., M.M.M., E.A.S.; managed project: S.L., D.M.M., R.R.H.A., R.A.G., A.B., J.S.J., E.A.S., B.D., S.R., T.F.C.M.; wrote and prepared manuscript: T.F.C.M., W.H., A.M., J.P., M.R., A.M.T., T.Z., A.B., J.S.J., E.A.S., S.R., B.D.

## References

- Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**: 974–984.
- Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–376.
- Andolfatto P, Depaulis F, Navarro A. 2001. Inversion polymorphisms and nucleotide variability in *Drosophila*. *Genet Res* **77**: 1–8.
- Assis R, Kondrashov AS. 2012. A strong deletion bias in nonallelic gene conversion. *PLoS Genet* **8**: e1002508.
- Astle W, Balding DJ. 2009. Population structure and cryptic relatedness in genetic association studies. *Stat Sci* **24**: 451–471.
- Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, Rollmann SM, Duncan LH, Lawrence F, Anholt RRH, et al. 2009. Systems genetics of complex traits in *Drosophila melanogaster*. *Nat Genet* **41**: 299–307.
- Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**: 519–520.
- Bellen HJ, Levis RW, Liao G, He Y, Carlson JW, Tsang G, Evans-Holm M, Hiesinger PR, Schulze KL, Rubin GM, et al. 2004. The BDGP gene disruption project: single transposon insertions associated with 40% of *Drosophila* genes. *Genetics* **167**: 761–781.
- Bellen HJ, Levis RW, He Y, Carlson JW, Evans-Holm M, Bae E, Kim J, Metaxakis A, Savakis C, Schulze KL, et al. 2011. The *Drosophila* gene disruption project: progress using transposons with distinctive site specificities. *Genetics* **188**: 731–743.
- Braig HR, Zhou W, Dobson SL, O'Neill SL. 1998. Cloning and characterization of a gene encoding the major surface protein of the bacterial endosymbiont *Wolbachia pipiensis*. *J Bacteriol* **180**: 2373–2378.
- Bridges CB. 1935. Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster*. *J Hered* **26**: 60–64.
- Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.
- Cingolani P, Platts A, Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w<sup>1118</sup>; iso-2; iso-3. *Fly (Austin)* **6**: 80–92.
- Cameron JM, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genet* **8**: e1002905.
- Corbett-Detig RB, Hartl DL. 2012. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genet* **8**: e1003056.
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.
- Dobzhansky T. 1937. *Genetics and the origin of species*. Columbia University Press, New York.
- Drosophila* 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203–218.
- Dunning Hotopp JC, Clark ME, Oliveira DC, Foster JM, Fischer P, Muñoz Torres MC, Giebel JD, Kumar N, Ishmael N, Wang S, et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**: 1753–1756.
- Endelman JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* **4**: 250–255.
- Ellis LL, Huang W, Quinn AM, Ahuja A, Alfrejd B, Gomez FE, Hjelmsten CE, Moore CL, Mackay TFC, Johnston JS, et al. 2014. Intrapopulation genome size variation in *D. melanogaster* reflects life history variation and plasticity. *PLoS Genet* (in press).
- Falconer DS, Mackay TFC. 1996. *Introduction to quantitative genetics*, 4th ed. Addison Wesley Longman, Harlow, United Kingdom.
- Fillon GJ, van Bommel JG, Braunschweig U, Talhout W, Klind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**: 212–224.
- Flint J, Mackay TFC. 2009. Genetic architecture of quantitative traits in mice, flies and humans. *Genome Res* **19**: 723–733.
- Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**: W273–W279.
- Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsøe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al. 2010. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**: 419–423.
- Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* **70**: 42–54.
- Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**: 269–276.
- Harbison ST, McCoy LJ, Mackay TFC. 2013. Genome-wide association study of sleep in *Drosophila melanogaster*. *BMC Genomics* **14**: 281.
- Hare EE, Johnston JS. 2011. Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods Mol Biol* **772**: 3–12.
- Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* **8**: 269–294.
- Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**: 756–766.
- Hoffmann AA, Rieseberg LH. 2008. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu Rev Ecol Evol Syst* **39**: 21–42.
- Hoffmann A, Turelli M, Simmons GM. 1986. Unidirectional incompatibility between populations of *Drosophila simulans*. *Evolution* **40**: 692–701.
- Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Res* **23**: 89–98.
- Huang W, Richards S, Carbone MA, Zhu D, Anholt RRH, Ayroles JF, Duncan L, Jordan KW, Lawrence F, Magwire MM, et al. 2012. Epistasis dominates the genetic architecture of *Drosophila* quantitative traits. *Proc Natl Acad Sci* **109**: 15553–15559.
- Jordan KW, Craver KL, Magwire MM, Cubilla CE, Mackay TFC, Anholt RRH. 2012. Genome wide association for sensitivity to chronic oxidative stress in *Drosophila melanogaster*. *PLoS ONE* **7**: e38722.
- Jovelin R, Cutter AD. 2013. Fine-scale signatures of molecular evolution reconcile models of indel-associated mutation. *Genome Biol Evol* **5**: 978–986.
- Kidwell MG. 1993. Lateral transfer in natural populations of eukaryotes. *Annu Rev Genet* **27**: 236–256.
- Kirkpatrick M, Barton N. 2006. Chromosome inversions, local adaptation and speciation. *Genetics* **173**: 419–434.
- Lee S, Ermond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**: 224–237.
- Leushkin EV, Sutormin RA, Nabieva ER, Penin AA, Kondrashov AS, Logacheva MD. 2013. Strong mutational bias toward deletions in the *Drosophila melanogaster* genome is compensated by selection. *BMC Genomics* **14**: 476.
- Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589–595.
- Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.
- Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**: 833–835.
- Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q, Vilhjálmsson BJ, Korte A, Nizhynska V, et al. 2013. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat Genet* **45**: 884–890.
- Mackay TFC. 2014. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet* **15**: 22–33.
- Mackay TFC, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet* **10**: 565–577.
- Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**: 173–178.
- Madsen BO, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**: e1000384.

- Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond JR, Strelets VB, Wilson RJ, the FlyBase consortium. 2013. FlyBase: improvements to the bibliography. *Nucleic Acids Res* **41**: D751–D757.
- Massouras A, Hens K, Gubelmann C, Uplekar S, Decouttere F, Rougemont J, Cole ST, Deplancke B. 2010. Primer-initiated sequence synthesis to detect and assemble structural variants. *Nat Methods* **7**: 485–486.
- Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W, Ayroles JF, Dermitzakis ET, Stone EA, Jensen JD, Mackay TFC, et al. 2012. Genomic variation and its impact on gene expression in *Drosophila melanogaster*. *PLoS Genet* **8**: e1003055.
- McDonald MJ, Wang W-C, Huang H-D, Leu J-Y. 2011. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol* **9**: e1000622.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.
- McQuilton P, St Pierre S, Thurmond J, FlyBase Consortium. 2012. FlyBase 101—the basics of navigating FlyBase. *Nucleic Acids Res* **40**: D706–D714.
- Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. 2010. Detecting copy number variation with mated short reads. *Genome Res* **20**: 1613–1622.
- Mettler LE, Voelker RA, Mukai T. 1977. Inversion clines in populations of *Drosophila melanogaster*. *Genetics* **87**: 169–176.
- Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59–65.
- The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**: 1787–1797.
- Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. 2013. The origin, evolution and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* **23**: 749–761.
- Navarro A, Betran E, Barbadilla A, Ruiz A. 1997. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics* **146**: 695–709.
- Navarro A, Barbadilla A, Ruiz A. 2000. Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in *Drosophila*. *Genetics* **155**: 685–698.
- Nei M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.
- Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Stricker C, Gianola D, Schlather M, Mackay TFC, Simianer H. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet* **8**: e1002685.
- Ometto L, Stephan W, De Lorenzo D. 2005. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* **169**: 1521–1527.
- Onishi-Seebacher M, Korb J. 2011. Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. *Bioessays* **33**: 840–850.
- Petrov DA. 2002. Mutational equilibrium model of genome size evolution. *Theor Popul Biol* **61**: 533–546.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559–575.
- R Development Core Team. 2013. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korb J. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333–i339.
- Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.
- Richardson MF, Weinert LA, Welch JJ, Linheiro RS, Magwire MM, Jiggins FM, Bergman CM. 2012. Population genomics of the *Wolbachia* endosymbiont in *Drosophila melanogaster*. *PLoS Genet* **8**: e1003129.
- She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**: 927–930.
- Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, et al. 2010. A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* **20**: 273–280.
- Stalker HD. 1976. Chromosome studies in wild populations of *Drosophila melanogaster*. *Genetics* **82**: 323–347.
- Stone EA. 2012. Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines. *Genome Res* **22**: 966–974.
- Swarup S, Huang W, Mackay TFC, Anholt RRH. 2013. Analysis of natural variation reveals neurogenetic networks for *Drosophila* olfactory behavior. *Proc Natl Acad Sci* **110**: 1017–1022.
- Teixeira I, Ferreira A, Ashburner M. 2008. The bacterial symbiont *Wolbachia* induces resistance to RNA viral infections in *Drosophila melanogaster*. *PLoS Biol* **6**: e2.
- Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen J-Q. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**: 105–108.
- Turelli M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's  $\zeta$  meets the abdominal bristle. *Theor Popul Biol* **25**: 138–193.
- Van Raden PM. 2008. Efficient methods to compute genomic predictions. *J Dairy Sci* **91**: 4414–4423.
- van Steensel B. 2011. Chromatin: constructing the big picture. *EMBO J* **30**: 1885–1895.
- Waszak SM, Hasin Y, Zichner T, Olender T, Keydar I, Khen M, Stütz AM, Schlattl A, Lancet D, Korb J. 2010. Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS Comput Biol* **6**: e1000988.
- Weber AL, Khan GE, Magwire MM, Tabor CL, Mackay TFC, Anholt RRH. 2012. Genome-wide association for oxidative stress resistance in *Drosophila melanogaster*. *PLoS ONE* **7**: e34745.
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* **86**: 929–942.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**: 82–93.
- Yamamoto A, Anholt RRH, Mackay TFC. 2009. Epistatic interactions attenuate mutations affecting startle behaviour in *Drosophila melanogaster*. *Genet Res* **91**: 373–382.
- Yang W, Woodgate R. 2007. What a difference a decade makes: insights into translesion DNA synthesis. *Proc Natl Acad Sci* **104**: 15591–15598.
- Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865–2871.
- Zichner T, Garfield DA, Rausch T, Stütz AM, Cannavó E, Braun M, Furlong EE, Korb J. 2013. Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res* **23**: 568–579.

Received December 20, 2013; accepted in revised form April 1, 2014.



---

## Acknowledgements

Gracias al Dr. Antonio Barbadilla, primero, por darme la oportunidad de poder hacer ciencia real por primera vez en mi vida. Tu manera de enseñar, aunque no se me note el 'estilo Barbadilla', I tengo dentro y muy en cuenta cada vez que debo transmitir. Agradecer además toda su paciencia en esta larga fase final de tesis.

Gràcies a la Dra. Sònia Casillas i a l'Emilio Centeno, els companys que em van donar la benvinguda al laboratori i em van acompanyar durant la primera etapa d'aquest viatge, deixant-me absorbir un munt de coneixements i habilitats noves.

Maite i David, gràcies pel treball en equip durant tot aquest temps, quan 3 persons tant diferents ens hem acabat complementant infinitat de vegades, i hem pogut avançar tots quan tot era novetat. Per compartir molt més que ciència, juntament també amb la Núria i la Yolanda. Hem acabat separats només físicament.

Agraïment a tots els membres (antics i actuals) i responsables dels altres laboratoris (tant actuals com ja no membres) del grup de Genòmica i Evolució, pel suport i bon ambient constant, tant discutint ciència com discutint del món més informalment davant unes pizzes cada divendres.

Moltes, moltes gràcies pare, mare i Marc, per un suport més que incondicional en tot els aspectes. Extendre-ho a tota la resta dels de casa, que m'heu empès tot i ser una cosa nova també per vosaltres.

Moltes gràcies a tothom que us heu preocupat per a mi, en especial als que heu acabat fent sempre la única pregunta tabú a un candidat a doctor: ¿i quan acabes?. Espero poder respondre diferent a la propera.

Thank you very much to everyone, a bit far away, that send me support, help and cheers during this adventure.

Special thanks to 'Two Steps From Hell', as a bannerman of the epic music genre. Without your music, my hours staring at sentences, thinking if those made sense at all, and slowly adding paragraphs, wouldn't been as productive and surely would have felt extremely longer.