Universitat Autònoma de Barcelona

Mapping natural selection through the Drosophila melanogaster development following a multiomics data integration approach

Marta Coronado Zamora

DOCTORAL THESIS

Mapping natural selection through the *Drosophila melanogaster* development following a multiomics data integration approach

MARTA CORONADO ZAMORA

Directors

Antonio Barbadilla Prados
Isaac Salazar Ciudad

Departament de Genètica i de Microbiologia
Facultat de Biociències
Universitat Autònoma de Barcelona
2018

ABSTRACT

Charles Darwin's theory of evolution proposes that the adaptations of organisms arise because of the process of natural selection. Natural selection leaves a characteristic footprint on the patterns of genetic variation that can be detected by means of statistical methods of genomic analysis. Today, we can infer the action of natural selection in a genome and even quantify what proportion of the incorporated genetic variants in the populations are adaptive. The genomic era has led to the paradoxical situation in which much more evidence of selection is available on the genome than on the phenotype of the organism, the primary target of natural selection.

The advent of next-generation sequencing (NGS) technologies is providing a vast amount of -omics data, especially increasing the breadth of available developmental transcriptomic series. In contrast to the genome of an organism, the transcriptome is a phenotype that varies during the lifetime and across different body parts. Studying a developmental transcriptome from a population genomic and spatio-temporal perspective is a promising approach to understand the genetic and developmental basis of the phenotypic change.

This thesis is an integrative population genomics and evolutionary biology project following a bioinformatic approach. It is performed in three sequential steps: (i) the comparison of different variations of the McDonald and Kreitman test (MKT), a method to detect recurrent positive selection on coding sequences at the molecular level, using empirical data from a North American population of *D. melanogaster* and simulated data, (ii) the inference of the genome features correlated with the evolutionary rate of protein-coding genes, and (iii) the integration of patterns of genomic variation with annotations of large sets of spatio-temporal developmental data (evo-dev-omics).

As a result of this approach, we have carried out two different studies integrating the patterns of genomic diversity with multiomics layers across developmental time and space. In the first study, we give a global perspective on how natural selection acts during the whole life cycle of *D. melanogaster*, assessing whether different regimes of selection act

through the developmental stages. In the second study, we draw an exhaustive map of selection acting on the complete embryo anatomy of *D. melanogaster*.

Taking all together, our results show that genes expressed in mid- and late-embryonic development stages exhibit the highest sequence conservation and the most complex structure: they are larger, consist of more exons and longer introns, encode a large number of isoforms and, on average, are highly expressed. Selective constraint is pervasive, particularly on the digestive and nervous systems. On the other hand, earlier stages of embryonic development are the most divergent, which seems to be due to the diminished efficiency of natural selection on maternal-effect genes. Additionally, genes expressed in these first stages have on average the shortest introns, probably due to the need for a rapid and efficient expression during the short cell cycles. Adaptation is found in the structures that also show evidence of adaptation in the adult, the immune and reproductive systems. Finally, genes that are expressed in one or a few different anatomical structures are younger and have higher rates of evolution, unlike genes that are expressed in all or almost all structures.

Population genomics is no longer a theoretical science, it has become an interdisciplinary field where bioinformatics, large functional -omics datasets, statistical and evolutionary models and emerging molecular techniques are all integrated to get a systemic view of the causes and consequences of evolution. The integration of population genomics with other phenotypic multiomics data is the necessary step to gain a global picture of how adaptation occurs in nature.

ACKNOWLEDGMENTS

Primero de todo quisiera agradecer muy sinceramente al Dr. Antonio Barbadilla y al Dr. Isaac Salazar por haberme dado la gran oportunidad de realizar este trabajo bajo su supervisión. Antonio, muchas gracias por haber guiado mis primeros pasos en la genética y haberme adentrado en este apasionante mundo. Isaac, gràcies per haver-me fet més crítica i haver contribuït en el meu creixement com a científica.

Quiero agradecer a David Castellano todo lo que me ha enseñado estos años. Te debo mucho, y esta tesis no habría sido posible sin ti. Así mismo, agradecer a Irepan Salvador toda su ayuda y haberme enseñado a ser más rigurosa. El trabajo con vosotros ha sido, sin duda, mucho más llevadero.

Agradecer a mis compañeros del laboratorio –Carla, Jon, Isaac, Sergi, Esteve y Jesús– todos los momentos compartidos. Carla, gracias por ofrecerme tu ayuda siempre que lo necesito y ser una gran amiga. Jon, a ti te debo los chismorreos de las tardes que tanto nos gustan y las risas que nos han hecho salir del laboratorio a coger aire. Isaac, tu hueco fue irremplazable y agradezco infinitamente tu apoyo en los momentos difíciles. Sergi, me alegro de que esta experiencia nos haya permitido conocernos mejor y haber aprendido tanto el uno del otro. Esteve, porque no todos los héroes llevan capa, siempre serás el mejor *sysadmin*. Jesús, mi gran compañero de fatigas, gracias por estar siempre ahí, envidio las ganas y la fuerza que le echas a todo.

Muchas gracias a todos los miembros, tanto antiguos como actuales, del grupo de GBBE. En especial a Sònia, Marta y Raquel, sois las culpables de que me enamorase de la bioinformática y la genómica durante la carrera. Alejandra, aprecio tus ánimos, las charlas y los buenos consejos que me has dado. Teresa, thanks for sending your cheers and support despite the distance. Marina, gracias por alegrar el cotarro las mañanas y tardes del café. Mario, gracias por saber transmitir lo que es hacer ciencia de verdad.

I thank Dr. Erich Bornberg-Bauer for giving me the opportunity to stay 3 months in his lab. Thanks to Carsten for his patience with me and his

follow-up guidance. I want to thank April and Brennen for making my days much more enjoyable.

També agrair a l'Alícia i a la Maria Josep pel seu ajut administratiu i haver-me facilitat tant les feines burocràtiques.

Agradecer a las fuentes de financiamiento de esta tesis: el Servicio de Genómica i Bioinformática del IBB por la ayuda durante los primeros meses y la Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya por la beca FI de 3 años.

Agradezco a mis amigas todos estos años que han permanecido a mi lado, aunque no nos hayamos visto todo lo que hubiéramos querido: Lorena, Lara, Dámaris, Eli, Sara, Tania y Vero. ¡Nos quedan muchas cenas, celebraciones y viajes por hacer!

A las personas que han sabido mantenerme cuerda en el mejor ambiente posible: Montse, Conchi y Teresa, gracias por enseñarme a relativizar los problemas. Después de bailar se ven las cosas de otra manera.

Gracias a Laia Blanes por su asesoramiento en el diseño de la portada.

Elias, from the beginning you were kind, supportive, and showed me that you were always there for me. Despite the distance, this feeling hasn't changed. *Ich liebe dich*.

Finalmente, agradecer a mi hermana y mis padres todo su apoyo incondicional, comprensión y ayuda que me han dado durante toda mi vida.

Muchas gracias a todos, tanto los que están lejos como los que ya no están, que me han apoyado, animado y cuidado durante este camino.

PUBLICATIONS

During the Ph.D. research, the following articles were published or are currently under submission process. Three of them are the core of the present thesis (articles 2, 3 and 4) while the others were published as the result of the collaboration with members of the Bioinformatics for Genomics Diversity (BGD) group at the UAB.

[1] Castellano, D., **M. Coronado-Zamora**, J. L. Campos, A. Barbadilla, and A. Eyre-Walker (2015). "Adaptive evolution is substantially impeded by Hill-Robertson interference in *Drosophila*." *Molecular Biology and Evolution* 33.2, pp. 442–455.

[2] Salvador-Martínez*, I., **M. Coronado-Zamora***, D. Castellano*, A. Barbadilla, and I. Salazar-Ciudad (2018). "Mapping selection within *Drosophila melanogaster* embryo's anatomy." *Molecular Biology and Evolution* 35.1, pp. 66–79.

[3] **Coronado-Zamora***, **M.**, I. Salvador-Martínez*, D. Castellano, A. Barbadilla, and I. Salazar-Ciudad. "Adaptation and selective constraint throughout *Drosophila melanogaster* life cycle." (Submitted).

[4] **Coronado-Zamora**, **M.**, J. Murga-Moreno, S. Hervás, S. Casillas, and A. Barbadilla. "Comparison of five McDonald and Kreitman test approaches using *Drosophila melanogaster* and human population data." (In preparation).

[5] Murga-Moreno*, J., **M. Coronado-Zamora***, A. Bodelón, A. Barbadilla, and S. Casillas (2018). "PopHumanScan: the collaborative catalog of human adaptation." *Nucleic Acid Research*.

* The authors contributed equally to the work.

Author contributions:

[1]: In this article, the evolutionary advantage of genetic recombination at the gene level was quantified for the first time in the *D. melanogaster* genome. I developed a pipeline to obtain genes related to the immune system and testis and performed a permutation test to demonstrate that these type of genes are under positive selection. This pipeline was later used in publication [3].

[2]: In this article, we have mapped by the first time phenotypic adaptation and natural selection over the complete anatomy of the embryo of *D. melanogaster*. Irepan Salvador-Martínez filtered the gene expression data, estimated the gene phylogenetic age and contributed to the design of the study. I performed the bulk of the analyses, calculated the selection estimators and estimated the genomic features. David Castellano estimated the polymorphism and divergence data and contributed to the design of the study.

[3]: In this article, the patterns of adaptation and constraint through the complete life cycle of *D. melanogaster* were analyzed. I performed the bulk of the analyses, calculated the selection estimators and estimated the genomic determinants. Irepan Salvador-Martínez performed the cluster analyses.

[4]: In this article, we have compared the performance of five McDonald and Kreitman test (MKT) approaches using *D. melangaster* and human data. I performed the bulk of the analyses, run simulations and contributed to the development of the R package and the iMKT web server. Jesús Murga-Moreno developed the core of the web server and contributed to the R package. Sergi Hervás developed the core of the R package and developed an extension of the asymptotic MKT approach.

[5]: In this article, we developed PopHumanScan, an online catalog listing and characterizing regions of the human genome that have been subjected to either recent sweeps or recurrent selection since the split between our species and chimpanzees. Jesús Murga performed the bulk of the analyses and contributed to the development of the database. I contributed to the design of the study and the development of the database.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

α – Proportion of substitutions fixed by adaptive evolution

d – Genetic distance between two orthologous sequences

D_N – Observed number of non-synonymous substitutions

d_N, K_a – Rate of non-synonymous substitutions per generation and site

D_S – Observed number of synonymous substitutions

d_S, K_s – Rate of synonymous substitutions per generation and site

F – Inbreeding coefficient

K – Evolutionary rate, fixation rate per generation and site

μ – Mutation rate per generation and site

m_N – Total non-synonymous sites

m_S – Total synonymous sites

N_e – Effective population size

T – Split time of two species

τ – Tissue specificity index

ω, d_N/d_S, K_a/K_s – Rate of non-synonymous substitution relative to the rate of synonymous substitution

θ – Watterson's estimator, a measure of nucleotide diversity

ω_a – Rate of adaptive non-synonymous substitution relative to the rate of synonymous substitution

ω_{na} – Rate of non-adaptive non-synonymous substitution relative to the rate of synonymous substitution

P_N – Observed number of non-synonymous polymorphisms

p_N – Rate of non-synonymous polymorphisms per non-synonymous site

P_S – Observed number of synonymous polymorphisms

p_S – Rate of synonymous polymorphisms per synonymous site

P_0 – Proportion of effectively neutral mutations

ANOVA – Analysis of variance

A/P – Anterior/Posterior

BDGP – Berkeley *Drosophila* Genome Project

BGS – Background selection

bp – Base pairs (nucleotides)

CDS – Coding region of a gene

cM – Centimorgan (unit for measuring genetic linkage)

CNS – Central nervous system

cv – Controlled vocabulary

DAF – Derived allele frequency

DFE – Distribution of fitness effects

DGN – *Drosophila* Genome Nexus

DGRP – *Drosophila* Genetic Reference Panel

DNA – Deoxyribonucleic acid

D/V – Dorsal/Ventral

eMKT – Extended MKT

FDR – False discovery rate

Fop – Frequency of optimum codons

FWW – Fay, Wyckoff and Wu MKT correction

GO – Gene ontology

HRi – Hill-Robertson interference

iMKT – Integrative MKT

MAF – Minor allele frequency

Mb – Megabase

MKT – McDonald and Kreitman test

ML – Maximum-likelihood

Mya – Million years ago

NI – Neutrality index

NGS – Next generation sequencing

ns – Not significant

PS – Phylostratum

RNA – Ribonucleic acid

RPKM – Reads per kilobase million

SD – Standard deviation

SFS – Site frequency spectrum

SNP – Single nucleotide polymorphism

TAI – Transcriptome age index

TDI – Transcriptome divergence index

UTR – Untranslated region

GLOSSARY

ADAPTATION – Any heritable trait that improves the ability of an individual organism to survive and to reproduce.

ADAPTIVE (POSITIVE, DARWINIAN) SELECTION – Selection for advantageous (beneficial) mutations, which leads to their fixation.

BACKGROUND SELECTION – Change in the frequency of neutral variants linked to deleterious sites.

BALANCING HYPOTHESIS – Hypothesis that proposes that most of the molecular variation observed in a population is maintained by balancing selection (Dobzhansky, 1970; Ford, 1971).

BALANCING SELECTION – Selection that maintains polymorphism at a locus.

BODY PLANT – Organism's morphology created by a reproducible spatial patterning of cells (Erwin, Valentine, and Jablonski, 1997).

BOTTLENECK – Drastic reduction in the population size leading to a greater influence of genetic drift.

CLASSICAL HYPOTHESIS – Hypothesis that proposes that purifying selection purges most of the genetic variation in a population (Muller and Kaplan, 1966).

CLONAL INTERFERENCE – In absence of recombination, the process by which beneficial mutations segregating simultaneously on different haplotypes compete for fixation.

CODON BIAS – A bias towards the use of one of the synonymous codons of a particular amino acid.

DEVELOPMENT – Process through which an embryo becomes an adult.

DEVELOPMENTAL BURDEN – The development of later stages is dependent on earlier stages so that high levels of conservation are expected to be found in the earlier stages of development (Riedl, 1978).

DIRECT SELECTION – A change in allele frequency that is caused by the effects on the fitness of the alleles which are being selected.

ECONOMY SELECTION HYPOTHESIS – Hypothesis that proposes that the energetic cost of transcription favors smaller introns in highly expressed genes (Castillo-Davis et al., 2002).

EXPRESSION BIAS – A measure of how much the expression of a gene is restricted to one or few developmental stages (or tissues).

EXPRESSION DIVERGENCE – A measure of how much the expression level of a gene has diverged over time.

FITNESS – The average contribution of a genotype to the next generation.

GENETIC DRIFT – The random change in allele frequency in finite populations over time (Kimura, 1968).

GENETIC DRAFT – Recurrent positive selection that leads to the reduction of linked neutral genetic diversity (Gillespie, 2000a,b, 2001).

GENOMIC DESIGN HYPOTHESIS – Hypothesis that proposes that larger introns are involved in a more complex regulation of the genes embedded on them (Eisenberg and Levanon, 2003).

GENOTYPE – The particular combination of alleles of a locus.

GERM LAYERS – Tissue layers formed during embryogenesis (ectoderm, mesoderm and endoderm).

HILL-ROBERTSON INTERFERENCE – The reduced efficacy of selection in regions of linkage disequilibrium due to the inference among selected alleles (Hill and Robertson, 1966).

HITCHHIKING – Change in the frequency of neutral variants linked to beneficial selected sites (Smith and Haigh, 1974).

INTRON DELAY HYPOTHESIS – Hypothesis that states that intron length plays a role in gene expression timing (Gubb, 1986).

LINKED SELECTION – Loss of neutral genetic diversity in the genome as an effect of selection at a linked site (see HITCHHIKING or BACKGROUND SELECTION).

MATERNAL-EFFECT GENE – Genes shed within the egg by the mother and never transcribed by the embryo.

MOLECULAR CLOCK HYPOTHESIS – Hypothesis that states that substitutions at a given protein occur at a constant rate over time (Zuckerkandl and Pauling, 1965).

NEUTRAL THEORY – The neutral theory states that the majority of variation within and between species is selectively neutral and is shaped by mutational input and genetic drift (Kimura, 1968).

PHENOTYPE – The observable traits of an organism.

PHYLOSTRATIGRAPHY – Statistical approach that consists on tracing the origin of a gene in a phylogenetic tree (Domazet-Lošo and Tautz, 2010).

PHYLOTYPIC STAGE – Most conserved stage (or period) of embryogenesis between species (Sander, 1983).

PLEIOTROPY – Phenomenon in which one allele affects more than one trait.

RUBY IN THE RUBBISH – In absence of recombination, the process by which beneficial mutations appearing on deleterious backgrounds are lost.

SELECTION COEFFICIENT – Measure of the relative fitness of a genotype.

SELECTIVE SWEEP – See HITCHHIKING.

SPATIAL PLEIOTOPY – Index measuring the pleiotropic effects of a gene on the embryonic anatomy.

SPATIAL TRANSCRIPTOMICS – Technique used to spatially analyze RNA-seq data, in individual cells or tissues.

TIME-COST HYPOTHESIS – Hypothesis that proposes that small introns are favored in genes that need to be transcribed quickly (Chen et al., 2005).

TRANSLATIONAL SELECTION – Selection for the efficient translation of a gene and folding of the protein product.

Chapter 1

INTRODUCTION

1 Introduction

1.1. Molecular population genetics

1.1.1. 50 years of molecular population genetics

The greatest innovation in the study of evolution at the genetic level along the past 50 years has been the use of molecular tools for studying the variation within and between species. Today, it is difficult to realize the revolutionary impact that the discovery of molecular variation has on the interpretation of genetic variation in natural populations (Casillas and Barbadilla, 2017; Charlesworth and Charlesworth, 2017).

The aim of molecular population genetics is to explain the genetic variation patterns at the population level from population genetic principles. The field was born in 1966, when two seminal articles provided the first measures of genetic variation in several allozyme loci. Half a century later, data has evolved from a bunch of sampled genes to the collection of thousands of complete genomes. Remarkably, the mathematical theory provided by the pioneering work of Fisher, Haldane, Wright and Kimura –i.e., dynamic models of allele frequency change in populations under the action of natural selection, genetic drift, mutation and/or gene flux–, still remains as the explanatory frame to account for the evolutionary change (Lynch, 2007; Charlesworth, 2010; Casillas and Barbadilla, 2017).

On the next pages, the main landmarks through the half-century of molecular population genetics are described, empathizing how the measurement of the genetic variation has been improved during that time, while in the theoretical side, the *neutral theory* of molecular evolution has become the universal null model to detect the footprint of natural selection in a genome.

From the allyzome era to the large-scale population genomics era

Three major advances in the data acquisition have been propitiated by successive molecular technologies: starting with the variation at the protein level, followed by the variation at the DNA sequence and ultimately the variation at the genomic level.

THE ALLOZYME ERA. The first empirical measures of genetic diversity were done in 1966 when the electrophoretically detectable variation (or allozymes) in both *Drosophila pseudoobscura* (Lewontin and Hubby, 1966) and humans (Harris, 1966) was described. These surveys of molecular diversity revealed an amount of variation much higher than it was expected from the two evolutionary selective views competing at that time (Lewontin, 1974). The so-called *classical hypothesis* (described by Dobzhansky in 1955 but attributed to Muller and Kaplan, 1966) stated that natural selection was the only force to purge variation, and therefore most loci were thought to be homozygous. Opposite to this hypothesis, the *balancing hypothesis* (Dobzhansky, 1955), postulated that a large proportion of loci were polymorphic and that the genetic variation was maintained by natural selection. Allozyme variability rejected the former hypothesis and seemed, at first, to support the latter one –but Kimura shortly showed that none of them could explain such variability, see section 1.1.2. The electrophoresis technique exhibits certain limitations, such as being unable to detect DNA variants which do not affect the mobility of a protein as well as variants which do not change the amino acid sequence (Lewontin, 1991). Therefore, allyzome data is not a sufficient source to measure genetic variation (Lewontin, 1991).

THE NUCLEOTIDE SEQUENCE ERA. The study of allozymes was replaced by a much more informative source of variation: the sequencing of nucleotide sequences, with the pioneering work of Kreitman in 1983. The availability of these sequences allowed the development of new statistical metrics to quantify variation (Casillas and Barbadilla, 2017). These statistical metrics are still actively used in numerous publications (reviewed by Vitti, Grossman, and Sabeti, 2013; Casillas and Barbadilla, 2017). Despite a large number of sequences for many genes and different species available (Clark et al., 2016) and the new tools developed to further characterize this diversity, these surveys could give a biased view of the genome variation patterns, as they were focused on providing the diversity in some particular regions of the genome.

THE POPULATION GENOMICS ERA. The first true population genomic work in *Drosophila* was done by Begun et al. in 2007, through the sequencing of six complete genomes of *Drosophila simulans* at low coverage. Today, a single run of an Illumina sequencing platform can provide more data than the one present in that study, facilitating the generation of big population genomics resources (Langley et al., 2012; Mackay et al., 2012; Lack et al., 2015). This revolution affects not only the population genomic data and the related new methodologies to deal with it but also other data associated with regulation, expression and other layers of the genome. All this data allows a better characterization of the targets of natural selection.

This thesis tries to face the challenges and opportunities of the present population genomics *momentum*. To understand the state of the art of population genomics, the conceptual basis of the neutral theory of molecular evolution first needs to be established.

1.1.2. The neutral and nearly neutral theories

Successfully defining the neutral theory requires the introduction of a key concept in population genetics: the *distribution of fitness effects* (DFE, Eyre-Walker and Keightley, 2007; Keightley and Eyre-Walker, 2010). Typically, new mutations that enter a population are categorized into three types based on natural selection in determining their fates: those that do not differentially affect the fitness of an individual carrying it (neutral), those that increase its fitness (beneficial), and those that decrease it (deleterious, or even lethal). However, there is, in reality, a continuous spectrum of fitness effects, from the most deleterious mutations to the most beneficial ones, defined in the density function DFE. The transformation of the underlying concepts of the DFE over time can be used to understand the history behind the rise of the theory of molecular evolution.

Molecular biology was born in the second half of the 20th century, at the time when molecular diversity was starting to be assessed with protein electrophoresis and protein amino acid sequencing methods. Pre-1960 ideas defended that all the differences within a species are due to mutations that are either deleterious (classical hypothesis, Figure 1.1A) or beneficial (balancing hypothesis, Figure 1.1B). In 1965, Emile Zuckerkandl and Linus Pauling proposed the *molecular clock hypothesis* (MCH) after

estimating that mammal hemoglobins evolve at a roughly constant rate of 1.4×10^{-7} amino acid substitutions per year. Motoo Kimura, in 1968, concerned by the unexpected big amount of variation present in the natural populations that neither balancing nor classical hypotheses could explain, and together with the previous observation that mutations accumulate linearly with time, proposed the radical hypothesis that most mutation changes in natural populations are neither harmful nor beneficial. Accordingly, *genetic drift* is responsible for the random fluctuations in allele frequency in finite populations. In the DFE, thus, appeared a new and predominant category of mutations: the neutral mutations (Figure 1.1C). A consequence of the neutral hypothesis is the minimal equation $K=\mu$ (Kimura, 1968). Under neutrality, the rate at which allelic changes are fixed in a given species (K) is equal to the mutation rate (μ). This simple equation underlies many tests to detect natural selection, such as the McDonald and Kreitman test (introduced in section 1.1.4). Also, this linear accumulation of substitutions over generations predicted by the neutral theory is the theoretical frame for the molecular clock hypothesis (generation-time effect).

Under the assumption of the molecular clock hypothesis, species with shorter generation times should evolve faster than those with longer generation times. However, this prediction of the neutral theory was challenged when it was discovered that rates of protein evolution were proportional to absolute time (in years), not to generation time. This means that protein molecular clocks of species with different generation time are similar, although the neutral theory predicts this constancy for species with an equal generation time (Zuckerkandl, 1976; Wilson, Carlson, and White, 1977). Tomoko Ohta redefined Kimura's neutral theory by introducing a new type of mutations: the nearly neutral mutations (Ohta, 1973). In the DFE, these mutations lie between the neutral and the deleterious ones, accounting for a significant fraction of all mutations (Figure 1.1D). The nearly neutral theory predicts that slightly deleterious mutations are mostly eliminated by natural selection in large populations, but a large fraction of them behaves as effectively neutral and are randomly fixed in small populations by genetic drift. This theory increasingly gained importance because it can be used as a robust null model against which to test any selective non-neutral hypothesis.

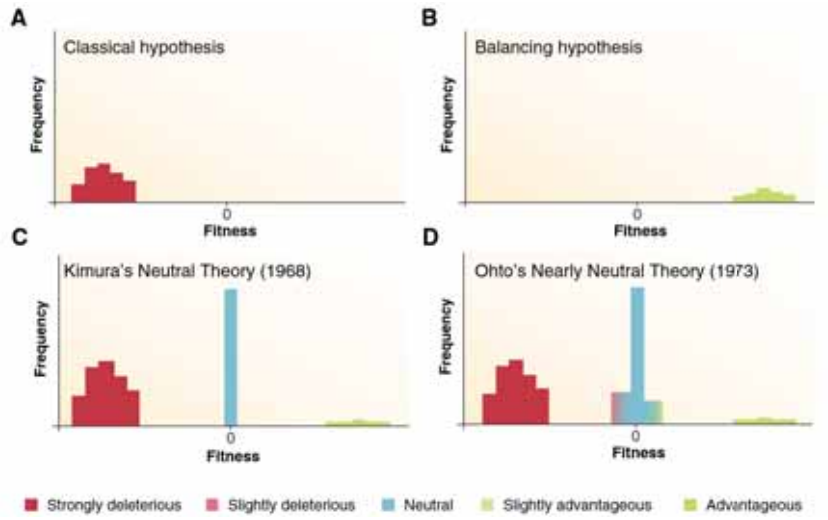


Figure 1.1 Historical evolution of the DFE. A. DFE representing the classical hypothesis (Dobzhansky, 1955; Muller and Kaplan, 1966). Natural selection is considered the only force to purge variation. **B**. DFE representing the balancing hypothesis (Dobzhansky, 1955). Natural selection maintains the genetic variation. **C**. In 1968, the DFE according to Kimura's neutral theory. The majority of differences within and between species are neutral. Some are adaptive (fixed by positive selection) and others are strongly deleterious. **D**. In 1973, the DFE according to Ohta's nearly neutral theory. The majority of differences are neutral, slightly deleterious or nearly neutral, and some are beneficial (fixed by positive selection) and others are deleterious. Figure adapted from Castellano (2016).

The molecular evolutionary rate as a function of the DFE

The DFE has been introduced as a key concept in population genetics and it is defined as a density function of the *fitness* (measured by the *selective coefficient*, s) of new mutations entering in the population, $f(s)$ (Figure 1.2A). However, the fate of new variation also depends on the probability of fixation of each new mutation. This probability depends on two factors: the strength of selection (s) and the *population size* (N) –assuming the simplification that the *effective population size* N_e equals N. Therefore, mutations with a selective coefficient s appearing in a population of size N have a probability of fixation defined as $u(N,s)$ (Figure 1.2B).

New variants enter in a population at a rate of $2N\mu$ (in a diploid population). Therefore, K, the *molecular evolutionary rate*, can be calculated as the integral of the combined probability of fixation and fitness effect

of all mutations that enter in a population, from fitness $-\infty$ to ∞ (also scaled to other intervals, e.g., from -1 to 1):

$$K = 2N\mu \int_{-\infty}^{\infty} u(N,s)f(s)ds \tag{1.1}$$

Considering the assumption of the neutral theory, which states that mutations are either neutral (s = 0) or strongly deleterious, the previous Equation 1.1 simplifies to:

$$K = 2N[\mu_0 u(N, s = 0) + (\mu - \mu_0)u(N, s = -\infty)] \tag{1.2}$$

The probability of fixation of a neutral mutation is $u(N,0) = 1/2N$, i.e., its initial frequency in the population, while the probability of fixation of a strongly deleterious mutation is zero, $u(N, s = -\infty) = 0$. Replacing these values in the previous Equation 1.2, the equation simplifies to Kimura's minimal equation $K = \mu$.

$$\begin{aligned} K = 2N\mu_0 u(N, s = 0) = 2N\mu_0 1/2N = \mu_0 \\ K = \mu_0 \end{aligned} \tag{1.3}$$

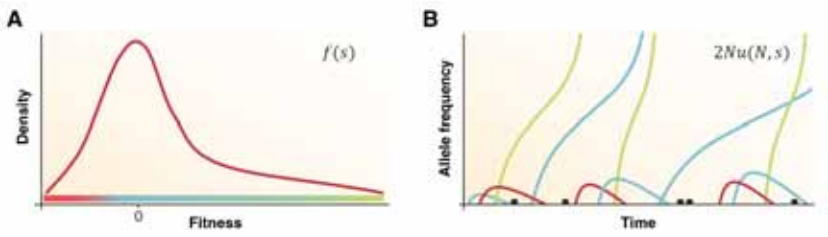


Figure 1.2 Factors determining the molecular evolutionary rate. A. Distribution of fitness effects (DFE) of new mutations. **B**. The probability of fixation of new mutations entering a population. The trajectory of mutations depends on the selective coefficient (*s*). New variants that enter in the population segregate over time and can become fixed or disappear. The colors represent the selective coefficients as in Figure 1.1.

The population size paradox

According to the (nearly) neutral theory, the average neutral nucleotide variation is the result of the equilibrium of two forces: mutation and genetic drift. Mutation adds genetic variation at a rate of $2N\mu$ (for diploid organisms) and genetic drift removes it at a rate that depends inversely on the population size ($1/2N$).

In small populations, genetic drift removes variation faster than mutation adds it, but in larger populations drift is less effective at removing variation. As a consequence, a relation between the effective population size and genetic variation is expected. The neutral heterozygosity parameter, θ, known as the Watterson's estimator (Watterson, 1975) predicts that small populations are expected to harbor less variation than large ones:

$$\theta = 4N\mu \tag{1.4}$$

However, this prediction was challenged with allozyme polymorphism data. While species' population sizes vary 20 orders of magnitude (Lynch, 2006), allozyme variation vary less than 4 (Bazin, Glémin, and Galtier, 2006). This observation is so-called Lewontin's paradox (Lewontin, 1974).

Smith and Haigh (1974) proposed an explanation to this paradox: the genetic *hitchhiking* effect (Figure 1.3A). Due to this process, neutral alleles that are linked to a favored selected mutation will also reach fixation –creating what was later called a *selective sweep* (Berry, Ajioka, and Kreitman, 1991). Genetic hitchhiking results in the reduction of the linked genetic variation, which could explain the observed genetic homogeneity of large populations. The levels of polymorphisms can be recovered over time by mutation and recombination.

This theory was revised when allozyme polymorphism data was replaced by DNA sequence data. This latter data showed low genetic variation in some regions of the genome, particularly near the centromeres or within chromosome rearrangements. The correlation between recombination and genetic variation was not found with divergence. This lack of correlation between recombination and divergence excludes the simplest explanation that recombination is mutagenic. This, in turn, leads to

the more feasible idea that the correlation between recombination and genetic variation is due to a greater elimination of the variation in regions of low recombination. John Gillespie developed a stochastic model which takes into account both the effects of genetic drift and recurrent hitchhiking, called *genetic draft* (Gillespie, 2000a,b, 2001). Genetic draft also removes genetic variation as genetic drift does, but the effect of genetic draft increases with the population size. In large populations, genetic drift is less effective at removing alleles and genetic variation increases. But at the same time, more adaptive mutations occur, because there are more alleles available to mutate and selection is more effective in larger populations. Therefore, genetic variation is reduced due to more hitchhiking events caused by the prevalence of adaptive selection. With this theoretical model, population size and genetic diversity can be uncoupled, potentially resolving Lewontin's paradox (Gillespie, 2004; Lynch, 2007).

Charlesworth et al. (1993) proposed a parallel effect regarding the deleterious mutations, known as *background selection* (BGS, Figure 1.3B). BGS is the process by which neutral variation is removed from the population for being linked to deleterious sites (Charlesworth, Morgan, and Charlesworth, 1993). While hitchhiking predominates when selection is strong and positive mutations are abundant, BGS will predominate when selection is relatively weak and mutations are recessive. BGS is expected to reduce the genetic variation similar to the hitchhiking effect. In this scenario, a neutral mutation can remain in the population if it appears in a chromosome free of deleterious mutations or if recombination breaks the haplotype. Therefore, recombination plays an important role in mediating the fate of linked sites in the genome (see section 1.1.3).

In 2015, Corbett-Detig estimated the expected reduction in neutral variation by both hitchhiking and BGS effects using empirical data from 40 eukaryotic species. This work showed that there is a positive correlation between the effective population size and the levels of linked neutral variation, i.e., natural selection removes more variation at linked neutral sites in species with large N_e than in those with small N_e. This work finally provided the necessary empirical evidence that natural selection constrains the levels of neutral genetic diversity across many species (Corbett-Detig, Hartl, and Sackton, 2015).

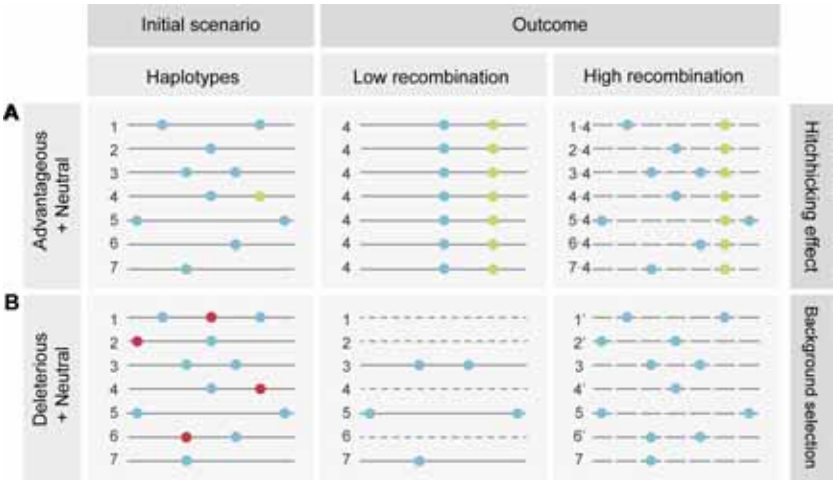


Figure 1.3 The effect of adaptive (hitchhiking) and purifying selection (BGS) on linked neutral sites in regions with low and high recombination. A. An adaptive mutation (green dot) destined to get fixed appears linked to a particular haplotype (i.e., linked to some particular neutral alleles, represented with blue dots). When the frequency of an advantageous allele increases in a region of low recombination, so does the frequency of the linked neutral variants and all the other neutral alleles are lost. This results in a reduction of the levels of polymorphism. In regions of high recombination, the neutral levels of polymorphisms can be recovered over time by mutation and recombination. **B**. BGS occurs when neutral alleles (blue) are linked to deleterious variants which will eventually be removed. In regions of low recombination, chromosomes carrying deleterious mutations (red dots) will be eliminated quickly from the population (represented as dashed gray lines). A neutral mutation can remain in the population if it appears in a chromosome free of deleterious mutations or if recombination breaks the haplotype. This results also in a reduction of the genetic variation not as strong as the hitchhiking effect. Figure adapted from Gómez-Graciani (2018) with permission.

1.1.3. The role of recombination in hitchhiking events

Recombination is a key parameter in modulating the patterns of genome variation (Casillas and Barbadilla, 2017). The fate of a new mutation appearing in a genome is conditioned not only by the selective advantage or disadvantage that confers to its bearer but also by the genomic context in which it appears. Population genetic theory predicts that an increased linkage between sites will limit the efficacy of both positive and purifying selection since selection at one site interferes with selection at linked sites (Hill and Robertson, 1966). If a newly selected mutation is surrounded by many other selected ones, these mutations will interfere with each other as they do not segregate independently. The reduction

of the efficacy of selection due to the interaction between linked sites is known as the *Hill-Robertson interference* (HRi, Hill and Robertson, 1966).

In low-recombining regions, there will be a greater density of selective alleles that do not segregate independently, and a lower efficacy of both positive and purifying selection is expected. On the other hand, high-recombining regions will exhibit higher adaptation rates due to a higher efficacy of both positive and purifying selection.

There are two kinds of HRi that can compromise the adaptation of genomes. In the first type, known as *clonal interference* (Figure 1.4A), adaptive mutations simultaneously segregating in different haplotypes compete for fixation (Gerrish and Lenski, 1998). Without recombination, they compete until the strongly beneficial one is fixed and the others are lost. A deleterious mutation linked to the beneficial mutation might be dragged to fixation as a consequence. The second type, called *Ruby in the rubbish* (Figure 1.4B), occurs when a beneficial mutation appears on genetic backgrounds loaded with segregating deleterious mutations (Peck, 1994). Without recombination, beneficial mutations in linkage with deleterious ones are lost. Both kinds of interference limit the rate of adaptation in genomes.

Castellano et al. (2015) quantified for the first time the evolutionary advantage of genetic recombination in the *Drosophila melanogaster* genome. Due to the HRi, the *D. melanogaster* genome has an adaptation rate around 27% below the optimal adaptation rate.

1.1.4. Detecting the footprint of natural selection

At the molecular level, natural selection can be divided into two types: *direct selection* refers to the variants that are the target of selection, and that can be either advantageous or deleterious; *linked selection* refers to the variants linked to those under selection, whether advantageous (hitchhiking, Smith and Haigh, 1974) or deleterious (BGS, Charlesworth, Morgan, and Charlesworth, 1993). Different methods have been developed to detect the action of natural selection at the molecular level. In this thesis, the effect of direct selection and the methods that have been used to detect it are addressed.

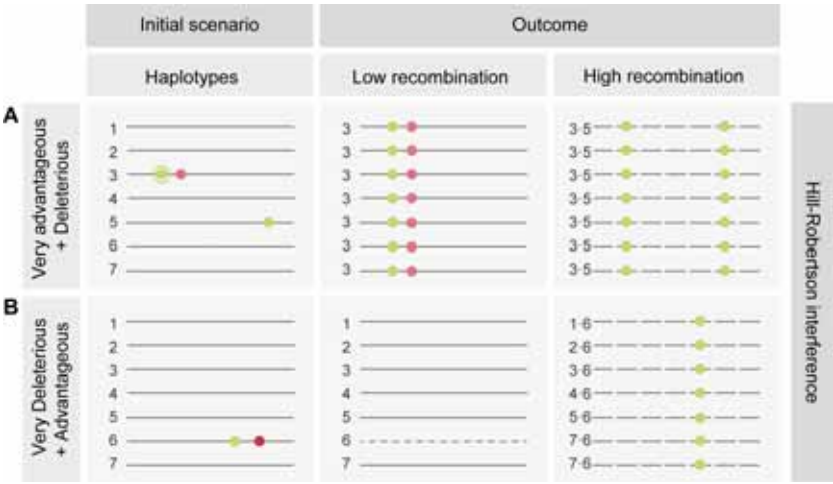


Figure 1.4 Two types of Hill-Robertson interference due to low recombination. **A**. Clonal interference. Is the process by which without recombination, beneficial mutations (green dots) segregating simultaneously on different haplotypes compete for fixation. A strongly beneficial mutation (green circled dot) is fixed together with a linked deleterious variant. **B**. Ruby in the rubbish. Is the process by which without recombination, beneficial mutations appearing on deleterious backgrounds are lost, also reducing adaptation. In both cases, only the haplotype(s) that rise(s) in frequency and may fix in the population. Figure adapted from Gómez-Graciani (2018) with permission.

The neutral theory provides the theoretical basis of the current tests of selection that are going to be presented in the next sections. Because the neutral theory makes clear predictions about the observed polymorphisms within species and fixed differences between species are neutral, it provides the necessary null (neutral) model against which to evaluate non-neutral hypotheses.

Detecting selection using divergence

The action of natural selection at the molecular level can be assessed by applying a simple method that does not require population genomic data but sequence data for differences between species. K, the evolutionary rate (introduced in section 1.1.2), is commonly estimated by the genetic distance between two orthologous sequences (d), divided by twice the divergence time of both species (T), because substitutions can occur on both branches of the phylogenetic tree:

$$K = d/2T \tag{1.5}$$

d is commonly estimated as the fraction of aligned amino acid positions that differ between the two sequences while correcting the effect of multiple substitutions (e.g., the Jukes and Cantor correction, 1969). This measure gives a general idea of how divergent two sequences are.

A powerful approach using divergence is to divide d depending on the type of mutation that it causes, either synonymous or non-synonymous. Therefore, d_N represents the number of replacement substitutions per non-synonymous site, while d_S is the number of silent substitutions per synonymous site. Since the mutation rate varies throughout a genome, Kimura (1977) suggested to correct d_N with d_S, which is equivalent to controlling for the differences in mutation rates, leading to the d_N/d_S ratio (also referred to as K_a/K_s ratio, or simply ω, Yang and Bielawski, 2000).

Although the d_N/d_S ratio includes the combined effect of neutral, advantageous and deleterious mutations, it can give an indication of the impact of natural selection on a sequence. The only accepted way to demonstrate that there have been advantageous mutations during the evolution of a sequence is to obtain a proportion higher than 1. A d_N/d_S ratio of 1 is obtained when all mutations are selectively equivalent. A ratio above 1 can only be obtained if there is a fraction of advantageous mutations. But this is a very stringent requirement, and only a few genes will ever reach a d_N/d_S higher than 1. One solution is not to account for the d_N/d_S of the overall protein (as the average will likely be smaller than 1) but subdividing the protein into functional domains, which may have a d_N/d_S higher than 1.

Another solution to increase the power of detecting positive selection by applying the d_N/d_S ratio is to analyze specific codons of a coding sequence in multiple species. This is implemented in the codon-based Z-test of the MEGA software (Kumar, Stecher, and Tamura, 2016).

The d_N/d_S test is a very conservative statistic for detecting adaptive evolution, but it can be informative about the selection pressure exerted on a sequence.

Detecting selection using polymorphism and divergence

The McDonald and Kreitman test (MKT) was born as a method to overcome the limitations of d_N/d_S when detecting adaptive selection (McDonald and Kreitman, 1991). This test takes into account both polymorphism and divergence data (Figure 1.5) and it is one of the most powerful methods to detect natural selection. The MKT normalizes the divergence ratio (D_N/D_S) with the polymorphism ratio (P_N/P_S) which takes into account the constraint at non-synonymous sites and thus increases the power of detecting adaptive selection.

Therefore, four different counts are needed to conduct the MKT: the count of polymorphisms at synonymous (P_S) and non-synonymous sites (P_N) as well as the count of fixed differences at synonymous (D_S) and non-synonymous sites (D_N). The four counts are placed in a 2×2 contingency table and significance is typically assessed with a Fisher's exact test. Note that in the d_N/d_S test the values used (d_N and d_S) are ratios since they are obtained as the total count of fixed differences (D) divided by the total number of sites in each type of site (m_N or m_S).

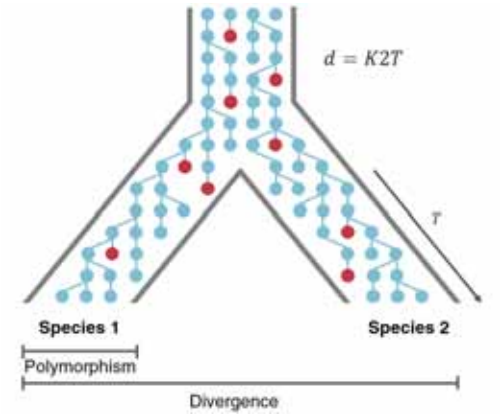


Figure 1.5 Diagram showing the trajectory of neutral mutations during a speciation process. Intrapopulation polymorphism is represented with blue dots which are segregating during the speciation process. These variants eventually get fixed or are lost (lost variants are represented with red dots). *T* is the divergence time; *K* the mutation rate and *d* the divergence rate.

Under neutrality, all non-synonymous mutations are expected to be neutral and the D_N/D_S ratio will roughly equal to the P_N/P_S ratio (Figure 1.6A). In contrast, positive mutations will rarely be detected as polymorphic variants, because they tend to be fixed quickly, but have a cumulative effect on the divergence. As a result, D_N/D_S will be greater than P_N/P_S (Figure 1.6B).

Some parameters have been developed to quantify selection using the MKT-test. One of these parameters is the neutrality index (NI, Rand and Kann, 1996).

$$NI = \frac{P_n / P_S}{D_N / D_S} \tag{1.6}$$

NI indicates to what extent the levels of polymorphic variation in the testing region depart from the expected level under the neutral model. Under neutrality, P_N / P_S equals D_N / D_S and thus NI equals 1. NI above 1 is interpreted as an excess of polymorphic variation compared to neutral regions which can be interpreted as due to purifying selection. NI below 1 can be interpreted as an excess of variation between species due to adaptive selection.

Another closely related summary statistic is the proportion of substitutions that have been fixed by adaptive evolution: α (Charlesworth, 1994; Smith and Eyre-Walker, 2002):

$$\alpha = 1 - \frac{D_S P_N}{D_N P_S} = 1 - NI \tag{1.7}$$

Defining α as a proportion is erroneous because it can also take negative values, and the meaning of a negative proportion is difficult to mathematically interpret. Positive values of α indicate an excess of divergence that is due to positive selection.

MKT is also used combining data across multiple genes by taking into account all the counts of polymorphism and divergence at synonymous and non-synonymous sites. However, this can lead to the Simpson's paradox (Simpson, 1951), wherein the results of individual 2×2 contingency tables suggest a trend which disappears or reverses when the data is combined in a single table. Regarding MK data, this can happen when there are large differences in the number of non-synonymous substitutions (D_N) between genes (Walsh and Lynch, 2018). To take into account this heterogeneity between genes, Stoletzki and Eyre-Walker (2011) suggested a weighted approach proposed originally by Tarone (1981) and Greenland (1982) which yields to unbiased estimated of the mean NI:

$$NI_{TG} = \frac{\sum D_{Si} P_{Ni} / (P_{Si} + D_{Si})}{\sum P_{Si} D_{Ni} / (P_{Si} + D_{Si})} \tag{1.8}$$

where the index refers to *i*th gene. With this index, each gene is weighted by its total synonymous variation (P_S and D_S).

MKT-based extensions

Assuming that neutral mutation rates at synonymous and non-synonymous sites are constant over time, adaptive selection is inferred as an excess of D_N/D_S ratio, and the D_N/D_S ratio is statistically higher than the P_N/P_S ratio. However, the MKT is not only statistically significant when there is an excess of divergence. The test can also be significant when there is an excess of non-synonymous polymorphism (i.e., P_N/P_S is significantly higher than D_N/D_S). The most common explanation for this excess is that slightly deleterious mutations contribute more to polymorphism than to divergence (because slightly deleterious mutations will not usually reach fixation), violating the assumption that all non-synonymous polymorphisms are neutral (Figure 1.6C). The presence of this non-neutral polymorphism can mask the effect of adaptive selection as it acts in opposite directions in the MKT (Figure 1.6D). The excess of non-synonymous polymorphism increases the P_N/P_S ratio, making the detection of adaptive selection unlikely due to a D_N/D_S lower than P_N/P_S.

Several methods try to deal with the presence of slightly deleterious mutations. Templeton (1996) suggested removing singletons (i.e., polymorphism only detectable in one sample) at both type of sites (synonymous and non-synonymous) from the total count of polymorphisms. The slightly deleterious substitutions, which segregate at low frequencies, should be overrepresented in the singletons (Figure 1.7C). Akashi (1999) proposed a more powerful method, that consisted of taking into account the complete distribution of allele frequencies, and not just the singletons. Although both extensions are more powerful than the MKT itself, their results are difficult to interpret, especially when the ratio of non-synonymous to synonymous differences varies among allele frequency classes (Hahn, 2018).

Fay, Wyckoff, and Wu (2001) developed a simpler methodology that removes all polymorphism at a frequency below a threshold (normally 5%–15%, Figure 1.7D). Although there is no consensus about what

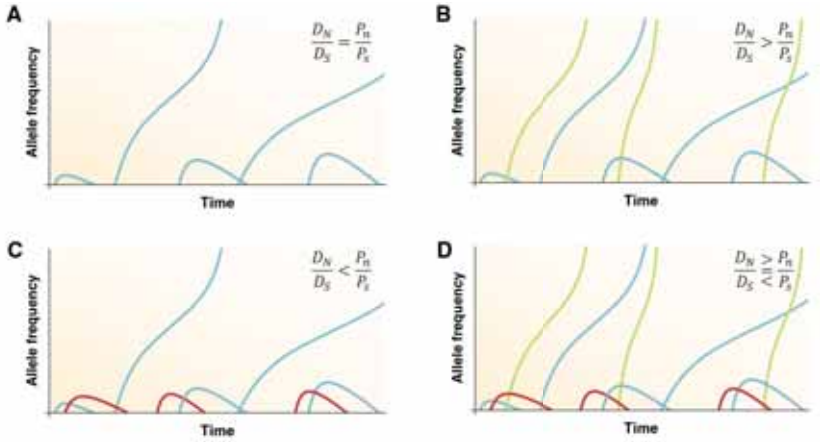


Figure 1.6 Different selective scenarios that can be inferred by the MKT. A. If only neutral alleles (blue) exist in the population, an equal proportion of divergent and polymorphic sites is expected ($D_N/D_S=P_N/P_S$). **B**. An excess of divergence compared to the polymorphism due to positive selection ($D_N/D_S>P_N/P_S$). **C**. If slightly deleterious substitutions segregate, they contribute to polymorphism ($D_N/D_S<P_N/P_S$). **D**. If both adaptive and deleterious variants are present, the result in the MKT can be misinterpreted. Adapted from Ràmia (2015).

this threshold should be exactly, Charlesworth and Eyre-Walker (2008) demonstrated that the α estimates are robust for a frequency threshold $\geq$15 (because slightly deleterious substitutions tend to segregate at low frequencies). However, this methodology is also expected to give biased α values, and only these estimates are reasonably accurate when the rate of adaptive evolution is high and the DFE of slightly deleterious mutations is leptokurtic (because leptokurtic distributions have a smaller proportion of polymorphisms that are slightly deleterious, Charlesworth and Eyre-Walker, 2008).

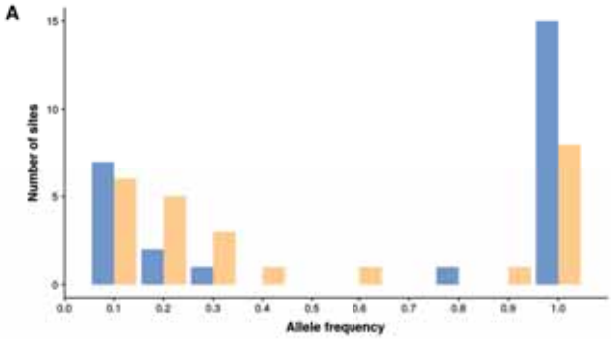
Mackay et al. (2012) proposed a powerful extension of the MKT called extended MKT (Figure 1.7E, eMKT method). Instead of simply removing low-frequency polymorphism below a given threshold, the count of segregating sites in non-synonymous sites is partitioned into the number of neutral variants and the number of weakly deleterious variants (see Methods, section 2.2.1, for methodological details). This increases the power of detecting selection and allows the independent estimation of both adaptive and weakly deleterious selection.

The last approach described here is the asymptotic MK method proposed by Messer and Petrov (2013). This extension is robust to the pres-

ence of selective sweeps (hitchhiking) and to the segregation of slightly deleterious substitutions (BGS). In this approach, α is defined as a function that depends on the site frequency spectrum (SFS) of the alleles, so α is estimated in different frequency intervals (x) and these values are then adjusted to an exponential function. An exponential fit is suitable as the non-synonymous allele frequency is expected to decay exponentially over the respective levels of synonymous polymorphisms (Messer and Petrov, 2013). The equation is:

$$\alpha = 1 - \frac{D_S P_{N(x)}}{D_N P_{S(x)}} \tag{1.9}$$

Figure 1.7 Comparison of different MKT approaches. Example of a gene exhibiting an excess of both slightly deleterious and fixed non-synonymous differences. **A**. The hypothetical allele frequency spectrum of synonymous and non-synonymous classes for a gene with *n*=10 sampled chromosomes. **B**. The standard MKT for this gene (*p*-value = 0.09, 2×2 Fisher's exact test). **C**. The 3×2 proposed by Templeton (1996), separating singleton polymorphism from all others (*p*-value = 0.07, 2×2 Fisher's exact test). **D**. The 2×2 table by Fay, Wyckoff, and Wu (2001), taking into account only polymorphism found on more than one chromosome (*p*-value = 0.045, 2×2 Fisher's exact test). **E**. Extended MKT (Mackay et al., 2012). The count of segregating sites in non-synonymous sites is partitioned into the number of neutral variants and the number of weakly deleterious variants. P_N is substituted with the number of non-synonymous polymorphism that is neutral (*p*-value = 0.042, 2×2 Fisher's exact test). Adapted from Hahn (2018).



B **Standard McDonald and Kreitman test (1991)**

	Polymorphism	Divergence
Non-synonymous	11	15
Synonymous	17	8

P-value = 0.092
2×2 Fisher's exact test

C **Templeton's correction (1999)**

	Singleton	Multiton	Divergence
Non-synonymous	7	4	15
Synonymous	6	11	8

P-value = 0.070
2×3 Fisher's exact test

D **Fay, Wyckoff and Wu's correction (2001)**

	Polymorphism	Divergence
Non-synonymous	4	15
Synonymous	11	8

P-value = 0.045
2×2 Fisher's exact test

E **Extended MKT (Mackay et al., 2012)**

	SFS ≤ 0.1	SFS > 0.1
Non-synonymous	7	4
Synonymous	6	11

$f_{neutral} = P_{s\leq0.1}/P_s = 6/17 = 0.35$

$P_{n\ neutrals0.1} = P_n \times f_{neutral} = 11 \times 0.35 = 3.88 \approx 4$

$P_{n\ neutral} = P_{n\ neutrals0.1} + P_n = 4 + 4 = 8$

	Polymorphism	Divergence
Non-synonymous	8	15
Synonymous	17	8

P-value = 0.042
2×2 Fisher's exact test

Figure 1.7 Comparison of different MKT approaches. Caption in the previous page.

MKT-extensions for other site classes

Although the estimation of the synonymous and non-synonymous counts is often done on coding sequences (as 0-fold and 4-fold sites are normally used as a proxy for non-synonymous and synonymous counts, respectively), it is possible to extend the MKT to other non-coding regions as long as one of the two sites are assumed to evolve neutrally (e.g., the case for intronic sites; Casillas, Barbadilla, and Bergman, 2007; Egea, Casillas, and Barbadilla, 2008). However, there are major caveats that have made it difficult up to this date to successfully apply the MKT to non-coding regions such as regulatory regions. Hahn (2018) raises consequentially the right questions when he asks: What is the genetic code of a regulatory region? How can the accumulation of changes positively affect a binding region? Those still unresolved questions need an in-depth inspection.

DFE-based approaches

The other family of MKT-derived methods is called DFE-based methods (Bustamante et al., 2005; Eyre-Walker and Keightley, 2007; Eyre-Walker and Keightley, 2009; Bustamante et al., 2002; Boyko et al., 2008; Eyre-Walker, 2006; Keightley and Eyre-Walker, 2007). These methods also correct the presence of slightly deleterious non-synonymous substitutions. The main distinction is that they first estimate the empirical DFE at the selected site class and then calculate how many non-adaptive substitutions are expected to become fixed given the inferred DFE from polymorphism data. For that, they usually also fit a demographic model to the synonymous dataset and assume all non-synonymous polymorphism are either neutral or deleterious. Simulations show that these methods can provide an accurate estimate of the average of α if data from a large number of genes are collected (Hahn, 2018). As sequencing becomes cheaper, single-locus studies of the DFE will soon become possible with such methods (Hahn, 2018).

One DFE-based method that is widely used is the DFE-alpha approach (Eyre-Walker and Keightley, 2009). DFE-alpha software incorporates the methodology for estimating the DFE of new deleterious mutations developed by Keightley and Eyre-Walker (2007) and the methodology for estimating α by Eyre-Walker and Keightley (2009).

For estimating α, this program infers the DFE of new deleterious mutations, and it uses this DFE to predict the number of substitutions orig-

inating from the neutral and slightly deleterious mutations. If the observed rate of substitutions (d_N) is greater than the expected rate, the excess of divergence can be attributed to the adaptive substitutions, yielding an estimate of α:

$$\alpha = \frac{d_N - d_S \int_0^\infty 2N\mu(N,s)f(s|a,b)ds}{d_N} \tag{1.10}$$

The method assumes that the DFE of deleterious mutations is a gamma distribution with scale parameter a and shape parameter b, $f(s|a,b)$. Given the inferred DFE, the average fixation probability of new deleterious and neutral non-synonymous mutations ($2N\mu(N,s)$) relative to the fixation probability of synonymous mutations is calculated by integrating over $f(s)$ (Eyre-Walker and Keightley, 2009).

Messer and Petrov (2013) conducted simulations to test the performance of their method, the asymptotic MKT, and DFE-alpha, in particular, to test their robustness to genetic draft or BGS. Simulations show that DFE-alpha is very robust and yields accurate α estimations even in the presence of genetic draft, BGS and demographic changes (Messer and Petrov, 2013). Furthermore, DFE-alpha does not require to set a minimum threshold for excluding the low-frequency polymorphisms (Eyre-Walker and Keightley, 2009).

1.1.5. Patterns of genome variation

Genetic diversity is ubiquitous. At the molecular level, the irrefutable proof is that no two humans share the same genome sequence (identical twins aside). With the advent of next-generation sequencing (NGS) technologies, we face data on million to billion variants. In fact, more than 6,000,000 variants have been described in the model species *D. melanogaster* up to date (Huang et al., 2014). What evolutionary forces could have led to such rich divergence between individuals within the same species? This unresolved question, that Gillespie (1991) referred to as the *great obsession* of population genetics, is reflected by a massive theoretical and empirical research trying to connect the patterns of variation in natural populations with the evolutionary forces that account for them (Connallon and Clark, 2014).

Known determinants of protein evolution

Understanding how do proteins evolve is a long-standing biological problem, and it is a central aim for evolutionary genetics. The large increase in the amount of available functional and molecular genomic datasets in the past few years provides an opportunity to shed light on this issue. In fact, factors affecting the rates of molecular evolution have long been studied, and many of them have been proposed to account for a great part of this variation, such as the expression level (Marais et al., 2004), expression bias (Duret and Mouchiroud, 2000), essentiality (Hirsh and Fraser, 2001) or codon usage (Holloway et al., 2008).

One of the major determinants of protein evolution seems to be the expression level of the gene coding for the protein (Pal et al., 2001; Drummond et al., 2005; Drummond, Raval, and Wilke, 2006), and this fact can be extended to a significant amount of organisms, both prokaryotes (bacteria, Rocha and Danchin, 2004) and eukaryotes (green algae, Popescu et al., 2006; *Drosophila*, Lemos et al., 2005; or *Arabidopsis thaliana*, Wright et al., 2004).

The expression level seems to account for approximately 30% of all variance of the protein evolutionary rate in all these organisms. These studies conclude that highly expressed protein-coding genes have a lower evolutionary rate (i.e., evolving slower) than those lowly expressed. Drummond and Wilke (2008) proposed that translation-induced protein misfolding may result in toxic products with detrimental effects. Under the assumption that non-synonymous mutations increase the probability of misfolding, the amino acid sequence of highly expressed proteins should thus be under stronger purifying selection than the sequence of proteins expressed at lower rates, irrespective of their biological function. On the other hand, in order to explain the observed low rate of synonymous substitutions in highly expressed genes, the translation accuracy hypothesis has been proposed (Akashi, 2003), which argues that since some codons are more favorable (adaptive) than others due to the biased distributions of tRNA, this could constraint synonymous change, too.

Thus, expression level accounts for a significant effect and must be considered as the main determinant to understand protein evolution. However, our knowledge of other features, such as the ones listed in Table 1.1, and their role as determinants of protein evolution has to be increased.

Table 1.1 Genomic features associated with protein evolutionary rates.

Feature	Correlation with ω	References
Exon length	Negative correlated	Guillén, Casillas, and Ruiz, 2018; Larracuente et al., 2008
Expression bias	Strong positive correlated, likely driven by positive selection on highly biased genes, and purifying selection on housekeeping genes	Guillén, Casillas, and Ruiz, 2018; Larracuente et al., 2008
Expression level	Strong negative correlated, likely driven by purifying selection	Rocha and Danchin, 2004; Lemos et al., 2005; Petit et al., 2007; Guillén, Casillas, and Ruiz, 2018
Essentiality	Essential genes are more conserved than nonessential genes	Hurst and Smith, 1999; Hirsh and Fraser, 2001
Dispensability	Weak association	Hirsh and Fraser, 2003
Functional category (ontology/-families)	Proteins implicated in multiple processes evolve slower. Extremely weak effect	Salathe et al., 2006
Intron length	Non-significant negative correlation	Comeron and Kreitman, 2000; Marais et al., 2005
Exon inclusion levels	Conflicting results: some indicate evidences of stronger positive selection in alternative regions, while others, the opposite	Sorek and Ast, 2003; Ermakova, Nurtdinov, and Gelfand, 2006
Intron number	Negative correlated	Marais et al., 2005
Mutation rate	Positive effect on protein evolution	Pál, Papp, and Lercher, 2006
Protein length	Weak negative correlated	Comeron, Kreitman, and Aguadé, 1999; Duret and Mouchiroud, 1999
Protein-protein interactions	Non-significant correlation	Giot et al., 2003
Recombination rate	Positive correlation with the efficacy of selection	Betancourt and Presgraves, 2002; Marais et al., 2004; Presgraves, 2005; Zhang and Parsch, 2005; Haddrill et al., 2007

Table adapted from Rocha (2006) and Larracuente et al. (2008).

1.1.6. The inquiry power of the population genomics approach

A complete population genomic study should be seen as a three-step process (Casillas and Barbadilla, 2017). The first step is to estimate the parameters that capture the evolutionary properties of the sequences

(e.g., polymorphism and divergence estimations or the proportion of adaptive fixations). These parameters are stored in population genomic browsers, such as PopFly (Hervas et al., 2017) or PopHuman (Casillas et al., 2018), which are the most complete population browsers for *Drosophila* and human populations respectively, up to date.

The second step, at the genomic level, is to correlate those population genomic parameters with other genomic variables throughout the genome. Such variables can be the recombination rate, gene density or GC density, to assess the impact of these features on the pattern of genetic variation.

The third step, the multiomics or integrative level, consists of the correlation between the patterns of genomic variation with annotations of large sets of omics data, e.g., transcriptomics. The main difference between -omics layers and the genomic sequence is that they change during the lifetime and body parts of organisms. The developmental transcriptome of an organism represents intermediate phenotypes between the genotype and the final phenotype on which natural selection acts (Civelek and Lusis, 2014). Although the expression is related to the genotype in a complex manner (determined by interactions between the genome and environment) the integration of -omics layers with population genomic data promises to provide a global view of the functional effects of genome variation (Wagner, 2008; Loewe, 2009; Pantalacci and Sémon, 2015; Casillas and Barbadilla, 2017).

1.2. *Drosophila* as a model organism

Fruit flies have been an attractive and effective genetic model since the Morgan laboratory at Columbia University worked on it to make crucial discoveries a century ago –proving the chromosome theory of inheritance (Morgan, 1910). Now, *D. melanogaster* is studied by more than 1,800 laboratories around the world (Hales et al., 2015): their easy culture, short generation time, many offspring, compact genome and easy manipulation make them a suitable organism for testing hypotheses in all research fields, including neurobiology, ecology, speciation, development, and of course population genetics (Powell, 1997). In the case of populations genetics, *D. melanogaster* has been crucial in three basic lines of research: chromosome inversion polymorphism evolution (Dobzhansky and Sturtevant, 1938), electrophoresis variation (Lewontin, 1974) and

nucleotide variation (Kreitman, 1983; Singh and Rhomberg, 1987). More recently, *D. melanogaster* has contributed to one of the major advances in the population genomic field providing the first high-resolution map of the footprint of natural selection (Mackay et al., 2012). This work demonstrates that natural selection is pervasive in *D. melanogaster* (further explored in section 1.4).

The genomic era has accelerated the development of high-throughput technologies. Today, more than 1,000 complete *Drosophila* genomes from multiple populations around the world are available (see section 1.4). Also, population data is becoming available for close *D. melanogaster* species, like *D. simulans* (Signor, New, and Nuzhdin, 2018). The availability of other non-model *Drosophila* species data (e.g., the i5k project, Thomas et al., 2018) is leading to powerful studies of comparative genomics.

1.2.1. Evolutionary history of *D. melanogaster*

There are more than 2,000 discovered *Drosophila* species in the world (Powell, 1997; Markow and O'Grady, 2006). One of the most extensively studied lineages is the *Sophophora* subgenus with around 330 species, including *D. melanogaster*, *D. simulans* and *D. yakuba* (Figure 1.8). These two latter species are commonly used as external species for divergence comparisons with *D. melanogaster* (outgroup).

D. melanogaster is currently accepted to be originated from Africa, and expanded to the rest of the world becoming a cosmopolitan species (Lachaise et al., 1988; David and Capy, 1988; Begun and Aquadro, 1993; Andolfatto, 2001; Stephan and Li, 2007). The expansion from Africa to Europe occurred 10,000–19,000 years ago (Li and Stephan, 2006; Thornton and Andolfatto, 2006; Duchen et al., 2013). *Drosophila* arrived in North America relatively recently, less than 200 years ago. The variation found in the non-African populations is lower than the one found in Africa (Begun and Aquadro, 1993; Andolfatto, 2001), indicating that the propagation of *D. melanogaster* outside Africa was preceded by a bottleneck (Begun and Aquadro, 1993; Andolfatto, 2001; Li and Stephan, 2006; Thornton and Andolfatto, 2006), which was deduced to be finished around 50 years ago (Thornton and Andolfatto, 2006; Karasov, Messer, and Petrov, 2010). A study by Duchen et al. (2013) suggested that the North American population is an admixture between the African and

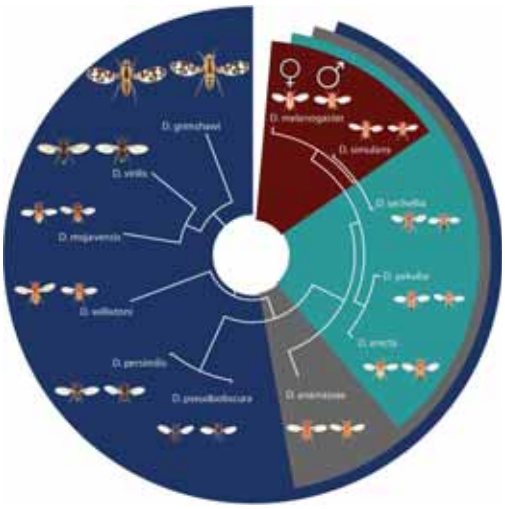


Figure 1.8 Phylogeny and taxonosmy of 12 *Drosophila* species. Species are organized into four major taxonomic groups indicated by color: *D. melanogaster* and *D. simulans* (n = 2, red), *melanogaster* subgroup (n = 5, light blue), *melanogaster* species group (n = 6, gray), 12 *Drosophila* genome species (n = 12, dark blue). Males (right) and females (left) of each species are presented and scaled according to their relative size. Figure taken from Stanley and Kulathinal (2016).

European populations. The cosmopolitan distribution makes *Drosophila* a very attractive organism to test how it has evolved and adapted independently to diverse environments (Schmidt et al., 2005; Markow and O'Grady, 2007).

1.2.2. The *D. melanogaster* life cycle

A major advantage of *D. melanogaster* is its particular short life cycle, which allows a high number of offspring to use in genetic crosses. Regarding its development, *D. melanogaster* is a holometabolous insect, which means it undergoes a complete metamorphosis –the immature stages (larva) are very different from the mature stages (adult).

On average, the complete cycle needs 9 to 10 days to complete in the lab at 25ºC (Figure 1.9). Upon fertilization, embryogenesis is completed in 24 hours, followed by three larval stages (first, second and third instar). Each instar lasts on average one day, except the third instar, which normally takes two days. Five days after fertilization, larval development is complete and it undergoes a complete metamorphosis in a pupal case, which takes 4–5 days. During this time, larval tissues break down and many adult structures develop from 19 different imaginal discs. Imaginal discs are a set of progenitor cells that give rise to the adult structures. Adult flies emerge from the pupal case (eclosion) and the process repeats

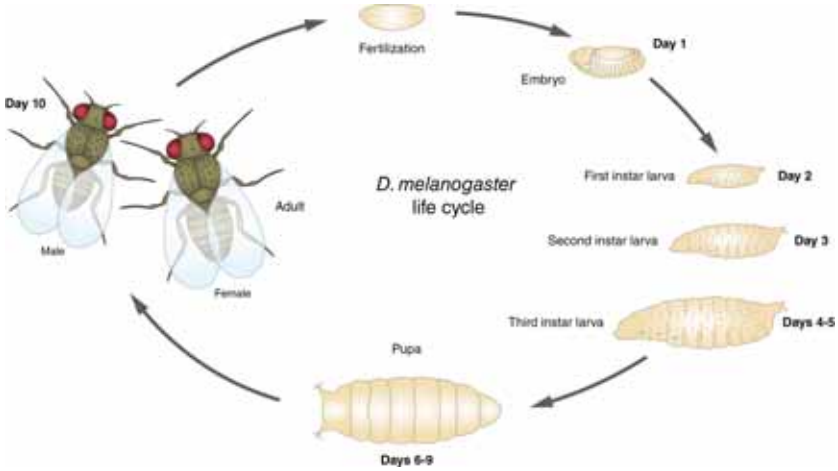


Figure 1.9 The *D. melanogaster* life cycle. The complete cycle takes 9–10 days when flies are maintained at 25°C in the lab. It is divided in four developmental stages: embryo (day 1), larva (days 2–5), pupa (days 6–9) and adult (day 10).

itself in 8–12 hours when the flies are sexually mature (Stocker and Gallant, 2008).

D. melanogaster embryo development

Embryo development is a continuous, complex process of interwoven temporal events. Some events are frequently emphasized in order to organize the development into a series of different stages. *D. melanogaster* embryo development has been divided into 17 stages by Volker Hartenstein and José Campos-Ortega (1993) and despite its artificiality, this division provides a useful temporal framework in which embryonic events can be referred to. A short description of the main events that characterize the *Drosophila* embryogenesis are described below (Figure 1.10).

FERTILIZATION. Embryogenesis starts with the fertilization of the oocyte (stage 1). Female flies store sperm for up to 2 weeks in specialized organs called seminal receptacles and spermathecae (Lefevre and Jonsson, 1962), which may contain sperm from different males. Because of that, sperm competition imposes important selective pressure on males. In fact, there is strong evidence that male reproductive genes evolve faster and are under adaptive evolution (reviewed by Swanson and Vacquier, 2002).

SUPERFICIAL CLEAVAGE AND CELLULARIZATION. After fertilization, the zygotic nuclei divide in a common cytoplasm with no new cellular membranes, referred to as syncytium (stage 2). After 10 synchronized rounds of division, nuclei migrate to the periphery, where they become partially encapsulated by cytoskeletal proteins to create furrow canals (stage 3/4). Cellularization occurs in stage 5, after the transcription of bulk zygotic genes, and marks the beginning of asynchronous cell divisions, followed by the gastrulation that takes places during stages 6 and 7 and is completed in stage 8.

GASTRULATION. The gastrulation determines the formation of the basic three germ layers: ectoderm, endoderm and mesoderm. Dramatic movements reshape the body plan. Cells from the posterior migrate towards the anterior in germ band extension (stages 9 and 10). From stage 11 onwards no major morphogenetic changes take place. The process follows with the germ band retraction to the posterior (stage 12). Cells then migrate to the dorsal midline in dorsal closure (stage 13), head structures begin to mature (stages 14–15), somatic musculature becomes visible (stage 16) and embryogenesis is nearly completed when the larval reaches its mature state (stage 17).

ANTERIOR/POSTERIOR PATTERNING. In *D. melanogaster*, the anterior-posterior (A/P) polarity is determined by the maternally contributed mRNA already present in the egg before fertilization. Translation of this mRNA after fertilization results in protein gradients, which are necessary for the specification of the expression patterns of a series of zygotic genes involved in segmentation and cell fate determination. There are three groups of segmentation genes (GAP genes, pair-rule genes and segment polarity genes) that are necessary for further specify A/P positioning within the embryo.

DORSAL/VENTRAL PATTERNING. Similar to the A/P patterning, the dorsal/ventral (D/V) is also determined through gradients of proteins, but the mechanisms are very different (Morisalo and Anderson, 1995). In the D/V patterning, the maternal genes already present in the oocyte are not transcribed until cellularization. Signaling events trigger a cascade leading to the embryonic patterning and axis determination.

MATERNAL-EFFECT GENES. The start of the embryogenesis involves mRNA and proteins already present in the egg, which are shed by the mother (King, 1970; Bastock and St Johnston, 2008). From a population genetic perspective, such genes are very interesting because they influ-

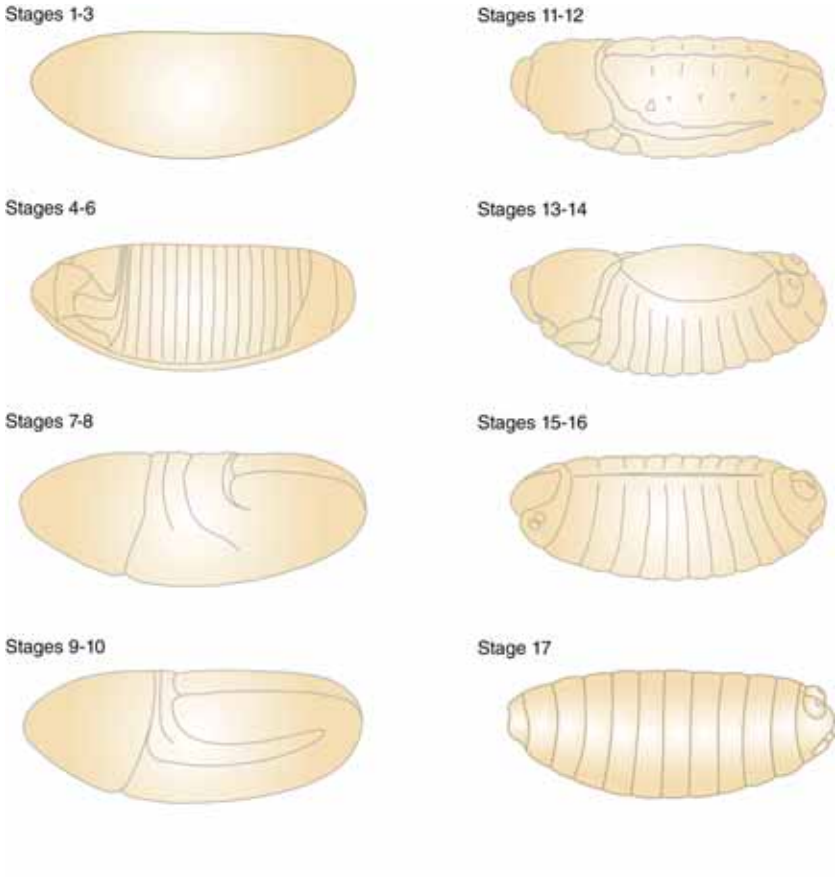


Figure 1.10 Main stages of *D. melanogaster* embryonic development. See section 1.2.2 for details. In each panel, anterior is to the right and dorsal is up. Images modified from the Atlas of *Drosophila* Development (Hartenstein, 1993) with permission.

ence the phenotype in the zygote from the mother's genotype, and not from the zygote's genotype. Therefore, the embryo of a homozygous mutant mother will be defective regardless of its own genotype. This is a so-called *maternal-effect* mutation. Selection on maternal genes will differ from selection on zygotic genes: selection is only half as strong when acting on a maternal-effect gene as it is when acting on a zygotic-effect gene. Because a maternal allele is not expressed in males, natural selection is relaxed by a factor of two relative to a zygotic allele, which is expressed in both sexes (Wade, Priest, and Cruickshank, 2009; Fairbanks, 2010).

1.2.3. Genome properties

D. melanogaster has a 180-megabase (Mb) genome organized in four different chromosome pairs: three autosome pairs labeled 2, 3 and 4 and the sexual pair. The sex chromosomes include an acrocentric X chromosome and a submetacentric Y chromosome, which is mainly composed of heterochromatin. In the case of autosomes, chromosomes 2 and 3 are large and metacentric and their arms are separately referred to as 2L, 2R, 3L and 3R. Chromosome 4 is very small, containing approximately 80 genes (Leung et al., 2010). Because of that, it is normally ignored in most studies (Hales et al., 2015).

D. melanogaster's small genome makes it a suitable model to use as a proof of concept for sequencing techniques and assembling of larger, more complex genomes (Rubin, 1996; Adams et al., 2000). In fact, *D. melanogaster* was the second metazoan genome to be sequenced after *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium, 1998) and the third eukaryotic genome after *Saccharomyces cerevisiae* (Goffeau et al., 1996).

The current number of annotated protein-coding genes in the *D. melanogaster* genome is 13,931 (according to FlyBase R.6.23; `http://flybase.org/cgi-bin/get_static_page.pl?file=release_notes.html`; last accessed: August 2018). Each protein-coding gene gets an annotation ID that begins with 2 letters ("CG"). Once protein-coding genes are further studied, they get a unique FlyBase identifier assigned, which begins with "FBgn".

1.3. Evo-devo: the link between genotype and phenotype

Development is the process through which an embryo becomes an adult. During this process, an organism's genotype is expressed to create the phenotype, on which natural selection primarily acts. The relationship between development (or ontogeny) and evolution (or phylogeny) has been a long debated topic since its origins (Darwin, 1872; Haeckel, 1879; Gould, 1977; Raff, 1996). The study of development in connection with evolution is important because changes in the adult morphology are first changes in the genes controlling the development leading to that

morphology (Alberch, 1980). Evolution cannot be understood without the consideration of developmental processes and how these affect evolution by affecting the phenotypic variation arising in each generation from genetic variation (Raff, 2000). The synthesis between developmental biology and evolution is within the scope of the field of evolutionary developmental biology, informally known as *evo-devo* (Gilbert, 2003).

1.3.1. Models of development

Despite the widely divergent final phenotype of all today's vertebrates, they go through a broadly similar appearance during embryo development. Different models of development have attempted to find a general relationship between animal development and evolution (Gould, 1977; Irie and Kuratani, 2014) which are summarized in the next sections.

The early conservation model

Karl Ernst von Baer first noticed in 1828 that there was a high similarity between animal species during periods of the embryogenesis. His observations on post-gastrulation embryos lead to the conclusion that early developmental stages are the most similar between species within a phylogenetic group, while late development stages are the most divergent ones (von Baer's third law, 1828, Figure 1.11).

Multiple theoretical justifications for Baer's third law have been proposed (Irie and Kuratani, 2011). The most plausible explanation suggests that changes in early development can have consequences in later development, since late developmental processes are causally dependent on the correct functioning of earlier developmental processes, while late developmental processes do not retroactively affect early developmental processes (Arthur, 1977; Riedl, 1978; Castillo-Davis and Hartl, 2002). As a result of this *developmental burden*, early development should be more constrained. Another possible explanation for this constraint is the *generative entrenchment* concept (Wimsatt, 1986), which similar to the developmental burden idea proposes that upstream regulators –i.e., any element that can affect gene expression– tend to be evolutionary conserved to ensure the correct generation of downstream events.

The hourglass model

The advent of developmental genetics in mouse and *Drosophila*, and some observations in comparative embryology (Sander, 1983; Elinson, 1987) led to an alternative hypothesis: the hourglass model of embryonic development evolution (Duboule, 1994; Raff, 1996, Figure 1.11). According to this hypothesis, early and late development would be more divergent between species than intermediate developmental stages (mid-development). The most conserved stage is called the phylotypic stage (Sander, 1983). Richardson (1995) proposed the term *phylotypic period*, given that there is not a unique conserved stage. There is no consensus about when the phylotypic stage appears. First, Ballard (1981) proposed that it appears in the pharyngula stage, Wolpert (1991) in the early somite segmentation stage and Slack, Holland, and Graham (1993) in the tailbud stage. Nowadays it is suggested that it should be in a stage directly after gastrulation, at least in arthropods (Wilt and Hake, 2004) or after neurulation in chordates (Wilt and Hake, 2004). There is also no consensus on what property is conserved. Some authors considered conservation of the expression patterns of specific genes (Haeckel, 1879; Duboule, 1994). For example, Duboule (1994) observed that the expression of the *Hox* genes –essential class of homeotic genes that specify the *body plans* of the developing embryo– are expressed during the phylotypic stage. The activation of these genes during the phylotypic stage will be responsible for the morphological conservation. On the other hand, Raff (1996) considered conservation of the developmental mechanisms. Raff (1996) proposed that the conservation was the result of complex interactions between developmental modules during the prototypic stage, which leads to selective constraints to reduce morphological divergence (Raff, 1996).

There have been proposed several hypotheses about the processes that may lead to an hourglass pattern. Some studies propose that many whole-body scale interactions take place during mid-development, while during early and late development interactions are at a much more restricted spatial scale (Raff, 1996), both at the level of mechanical interactions and molecular signaling between tissues. Accordingly, changes in mid-development would be much more likely to affect the whole embryo and changes at other stages would have much more spatially restricted effects. Other authors argue that the hourglass pattern arises from different selection pressures acting in early and late development

(Slack, Holland, and Graham, 1993; Wray, 2000; Wray, 2002; Kalinka and Tomancak, 2012).

Other developmental models

A number of alternative models explaining the patterns of embryonic conservation have been proposed. Richardson et al. (1997) proposed the adaptive penetrance model, which questions the existence of a phylotypic stage and contrarily proposes that the most beneficial mutations are likely to occur precisely in that stage (Richardson et al., 1997), because its potential to generate adult innovation (Figure 1.11).

Another model postulates no temporal difference of evolutionary conservation during development. This is the so-called ontogenetic adjacency model, proposed by Poe and Wake (2004, Figure 1.11).

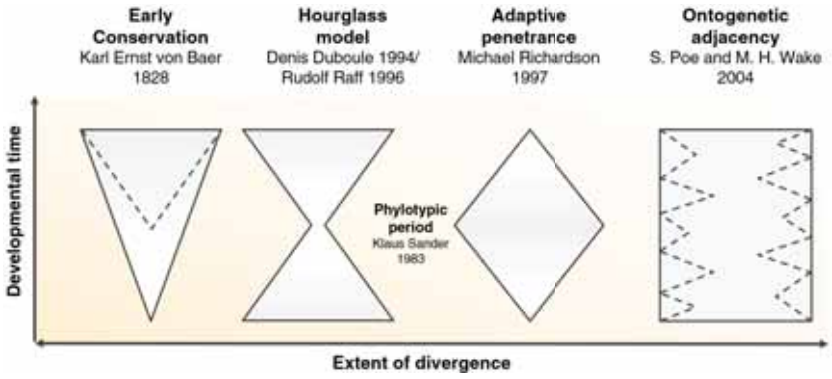


Figure 1.11 Models of development. Scheme showing four models for the embryo development. In all models, development from egg to adult is shown in the *y*-axis. Divergence is represented on the *x*-axis. In the *Early conservation model*, dashed lines indicate that von Baer based his observations on post-gastrulation embryos. In the *Ontogenetic adjacency model*, solid lines indicate that this model does not predict any particular pattern of conservation; dashed lines represent an instance of a conservation pattern. Figure from Kalinka and Tomancak (2012).

There is an ongoing discussion about whether the early conservation model, the hourglass model or some other pattern can explain the conservation among developmental stages (Richardson et al., 1997; Poe and Wake, 2004; Kalinka and Tomancak, 2012). In the next section, the different molecular approaches to test developmental models are presented.

1.3.2. Testing evo-devo models with molecular data

Studies assessing the mechanisms underlying the conservation of the development traditionally have used morphological approaches. However, the advent of molecular technologies has provided new molecular data to test developmental models.

To quantify the similarity between cross-species embryos using molecular data, three main approaches have been used: comparative genomics, comparative transcriptomics and phylotranscriptomics.

First, comparative genomics assesses the sequence divergence or conservation of orthologous genes expressed in different developmental stages to explore whether some stages are more conserved than others. Sequence conservation is typically measured using the d_N/d_S statistic (see section 1.1.4).

Second, comparative transcriptomics approaches, which has become a common method of choice in evo-devo (reviewed in Roux, Rosikiewicz, and Robinson-Rechavi, 2015; Pantalacci and Sémon, 2015). This approach consists of estimating the expression level for the whole transcriptome in several species, including time-course data of gene expression following organs or stages during their development. A classical use of this approach is to measure the *expression divergence* of orthologous genes in time-course data.

Third, phylotranscriptomics quantifies the evolutionary age of the developmental transcriptome (Domazet-Lošo and Tautz, 2010, reviewed in Drost et al., 2017). The difference in age of the genes expressed in different developmental stages has been suggested to be a good indicator of evolutionary conservation of the gene sequence (Irie and Sehara-Fujisawa, 2007; Domazet-Lošo and Tautz, 2010). Determining the age of a gene can be done following Domazet-Loso's approach (2007), which consists of tracing the origin of the gene in a phylogenetic tree using BLAST. Each gene can be assigned to a phylostratum, that represents the oldest phylogenetic node in which the gene can be found. Phylostratigraphic maps are already available for model species like *D. melanogaster*, *Danio rerio* (zebrafish) or *Arabidopsis thaliana* (Drost et al., 2015).

These three approaches independently showed that the conservation at the sequence, expression and evolutionary age level for orthologous genes seems to be maximal during the mid-development (Domazet-Lošo

and Tautz, 2010; Kalinka et al., 2010; Irie and Kuratani, 2011; Yanai et al., 2011; Levin et al., 2012; Wang et al., 2013; Gerstein et al., 2014; Drost et al., 2015; Levin et al., 2016). One of the most complete studies is the one from Levin et al. (2016). In this study, the developmental transcriptome series in ten species belonging to ten different phyla were compared in order to test the existence of a phylotypic period across phyla. An inverted hourglass pattern (i.e., early and late conservation and mid-development divergence) is reported comparing different phyla while within phyla an hourglass pattern is inferred.

Molecular studies in *D. melanogaster*

In this section, a short revision of the main contributions provided by the three main molecular approaches described above is presented for *D. melanogaster*.

COMPARATIVE GENOMICS. Davis, Brandman, and Petrov (2005) estimated the sequence conservation using the d_N/d_S statistic on 4,028 orthologous genes between *D. melanogaster* and *D. pseudoobscura*. By combining the d_N/d_S analysis with microarray expression data from different developmental stages, they found that genes with the highest rates of non-synonymous substitutions were expressed at low levels in late embryonic development and at high levels in the larva, pupa and adult. The genes with the lowest rates of non-synonymous substitution (the most conserved genes) were expressed at high levels in late embryonic development and at low levels before and after late embryonic development. This suggests, according to the authors, an hourglass pattern where embryonic stages spanning from 12 to 22 hours are highly conserved between *D. melanogaster* and *D. pseudoobscura*. A similar study by Mensch et al. (2013) estimated d_N/d_S for more than 2,000 genes across six different *Drosophila* species for three categories of genes: maternal genes, genes expressed in early development and genes expressed in late development. Maternal genes and late embryonic genes show higher d_N/d_S than early expressed genes. Finally, another study by Artieri, Haerty, and Singh (2009) found that genes expressed in the adult have a higher d_N/d_S than genes expressed in the pupa. The pupa, in turn, has a higher d_N/d_S that those expressed in the embryo (for 7,180 analyzed genes), thus favoring von Baer's law. Similar studies exist for other species of animals, zebrafish and mouse (Roux and Robinson-Rechavi, 2008; Pi-

asecka et al., 2013) and plants (*A. thaliana*, Quint et al., 2012; Gossmann et al., 2016) with similar conflicting results.

COMPARATIVE TRANSCRIPTOMICS. Kalinka et al. (2010) in a comparative transcriptomics approach, compared the genome-wide expression profiles across embryo development in six species of *Drosophila*. The time-course expression for 3,019 orthologous genes was measured by microarrays in eight 2-hour intervals during embryo development. In agreement with the hourglass model, the study found that mid-development, around the 10-hour stage, is the period in which gene expression levels is the most transcriptionally conserved among the six species.

PHYLOTRANSCRIPTOMICS. Domazet-Lošo and Tautz (2010) performed a study using a phylotranscriptomics approach. By overlapping the gene age classification on the developmental transcriptome of *D. melanogaster*, they show that genes expressed in the phylotypic stage are the oldest and most conserved, while genes expressed in early and late development are younger, also favoring the hourglass model.

1.3.3. The origin of the germ layers

An intriguing observation in animal development is the formation of the germ layers. All true animals have at least two germ layers: the ectoderm and the endoderm. The majority of animals, with the exception of diploblastic animals (Cnidaria and Ctenophora), has a third one, called the mesoderm. Despite their importance in development, there is no consensus theory on how two-layered metazoans appeared (reviewed in Technau and Scholz, 2003).

The origin of the germ layers was addressed in the nematode *C. elegans* by Hashimshony et al. (2015). Using single-cell RNA-seq data (Hashimshony et al., 2012), gene expression in each germ layer was determined. The different germ layers were found to have distinct global gene expression dynamics. The authors found sequential inductions of the genes expressed in each layer: first, endoderm genes are expressed, followed by ectoderm genes and finally by mesoderm genes. This suggests a recapitulation of the evolutionary appearance of the germ layers and accordingly, the endoderm genes are inferred to be the oldest ones, while mesoderm genes, the last ones to develop. Another evidence for this sequential origin of the germ layers was found by using phy-

lostratigraphy maps (Domazet-Lošo and Tautz, 2010), which concluded that endoderm genes are older than the ectoderm and mesoderm ones. However, in a study in *D. melanogaster* by Domazet-Loso, Brajković, and Tautz (2007), the ectoderm, and not the endoderm, was found to be the oldest germ layer. A plausible explanation for these contradictory observations is the quality of the data that was used: while in *C. elegans* the whole transcriptome was used, in *D. melanogaster* the analyses were performed on a limited number of tissues (Yanai, 2018).

1.3.4. Evo-devo in the genomics era (evo-dev-omics)

Some examples have been shown of how comparative embryology can benefit from the genomic era. In this section, the new techniques, derived from the advent in sequencing and transcriptomic technologies, are summarized (reviewed in Kalinka and Tomancak, 2012; Yanai, 2018, Figure 1.12).

From gene annotations and sequence alignments, microarrays and RNA-seq experiments can be designed. Microarrays (Schena et al., 1995) allow the measurement and comparison of genome-wide gene expression in multiple species. RNA-seq (Wang, Gerstein, and Snyder, 2009) allows a fine measurement of the expression levels of individual genes and cross-species comparisons of gene expression levels. Genes can then be clustered accordingly to their expression patterns during development.

Expression data can be integrated with other annotations, as for example gene set enrichment analyses and phylostratigraphy approaches. For a given transcriptome, it is useful to assess which functional categories of genes are enriched (for example, performing a gene ontology enrichment analysis). The age of genes can be inferred from the pattern of occurrences of orthologous genes in other genomes using phylostratigraphy maps, which can be used to assess the genes' sequence conservation.

Another interesting application of transcriptomic data lies at the integration of evo-devo and population genetics, with the main goal of finding specific loci associated with adaptation (Martin and Orgogozo, 2013; Pantalacci and Sémon, 2015). However, measuring adaptation directly on phenotypic traits is challenging and time-consuming, and therefore most studies on phenotypic adaptation are limited to a single or small

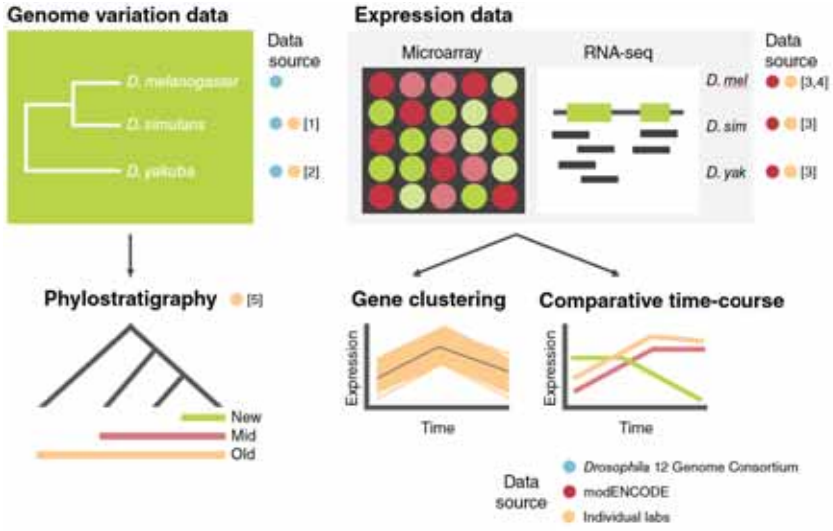


Figure 1.12 Evo-dev-omics approaches. Transcriptomics and phylostratigraphy tools for studying animal development. Complete genomic sequences are available for many related examples. Microarray and RNA-seq expression data can be used to determine developmental gene expression time-courses and perform gene clustering analysis. Phylostratigraphy approaches can be used to determine the gene age. Data sources: [1] Hu et al. (2013); [2] Rogers et al. (2014); [3] Bastian et al. (2008); [4] Arbeitman et al. (2002); [5] Drost et al. (2015).

number of traits per organisms, for example, the olfactory specialization in mosquitoes (Rinker et al., 2013) or the efficient osmoregulation in desert-adapted rodents (Marra, Romero, and DeWoody, 2014).

1.4. Population and evolutionary genomics in *Drosophila*

The *Drosophila* community has access to powerful resources for carrying out population and evolutionary biology analyses. The main contributions that have enabled advances at the intersections of population and evolutionary genomics are highlighted below.

With the development of high-throughput sequencing technologies, *D. melanogaster* has up to date more than 1,000 complete genomes from populations around the world (Lack et al., 2015, 2016, Figure 1.13). One of the most important contributions was provided by two independent,

but largely complementary projects: the *Drosophila* Genetic Reference Panel (DGRP, Mackay et al., 2012) and the *Drosophila* Population Genomics Project (DPGP, Langley et al., 2012). The DGRP is a panel of 205 complete genomes of *D. melanogaster* from a population sampled in Raleigh, North America (Mackay et al., 2012; Huang et al., 2014). In turn, the DPGP focused on the genomes of *D. melanogaster* from populations in Africa and France (Langley et al., 2012; Grenier et al., 2015; Lack et al., 2015).

The *Drosophila* Genome Nexus (DGN) is a recent compilation of each of these population genomic sequences, aligned using a common reference alignment pipeline, which facilitates direct comparison among datasets (Lack et al., 2015, 2016). Up to date, the DGN provides 1,121 wild-derived genomes spanning much of the *D. melanogaster* current geographic range (Lack et al., 2016). Such population datasets have allowed studies about the demography and migration of *D. melanogaster* (Pool et al., 2012), the bases of local adaptation (Langley et al., 2012), chromosomal inversion (Corbett-Detig and Hartl, 2012) or copy number variation (Langley et al., 2012). From a population genomics perspective, the DGRP and DPGP projects have provided the first high-resolution map of the footprint of natural selection. Both projects show that both adaptive and purifying selection are pervasive in the genome of *D. melanogaster*. More intriguing, they show that 30% to 50% of all fixed non-synonymous substitutions in *D. melanogaster* are caused by adaptive selection (Eyre-Walker, 2006; Mackay et al., 2012).

1.4.1. *D. melanogaster* resources

The complete description of the general properties of a genome is the result of large-scale collaborations and the modENCODE project is one of the best examples of it. The project was launched in 2007 with the aim of defining the functional elements of the *D. melanogaster* and *Caenorhabditis elegans* genome (Celniker et al., 2009). modENCODE has successfully provided the scientific community with a broad view of the genome-wide gene regulation and structure of the genome of *D. melanogaster* (The modENCODE Project Consortium 2010).

Figure 1.13 Available population genomics resources for four *Drosophila* species. Population genomics resources available for four *Drosophila* species. Dots represents sequenced populations, and the size of the dot is proportional to the number of individuals sequenced. For details, see Figure 1 from Casillas and Barbadilla (2017). Figures taken from Casillas and Barbadilla (2017) with permission.

A wide variety of online resources and other complementary efforts are available for *D. melanogaster* to the scientific community, contributing to the extensive and comprehensive characterization of the genome. The main databases of genomic resources, population and -omics datasets devoted to this species are compiled in Table 1.2.

Table 1.2 Main genomic, population and -omics resources available for *D. melanogaster*. Updated at the time of writing, August 2018.

Data type	Data source	Link (last update)	References
RNA expression	modENCODE	http://flybase.org (v.FB2018_04, August 23, 2018) http://www.modencode.org/	Graveley et al. (2011) Gelbart and Emmert (2013)
	Bgee	https://bgee.org (v.14.0, February 14, 2018)	Bastian et al. (2008)
RNA localization	FlyAtlas2	http://flyatlas.gla.ac.uk/FlyAtlas2/index.html (v.2, August 15, 2018)	Leader et al. (2018)
	FlyExpres	http://www.flyexpress.net/ (v.7, 2017)	Kumar et al. (2011) Kumar et al. (2017)
	Fly-FISH	http://fly-fish.ccbr.utoronto.ca/ (2016)	Lecuyer et al. (2007) Wilk et al. (2016)
	BDGP in situ	http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl (v.3, August 22, 2018)	Tomancak et al. (2002) Tomancak et al. (2007)
Regulatory regions	RedFly	http://redfly.ccr.buffalo.edu/ (v.5.4.3, July, 24 2018)	Gallo et al. 2011
Metasites	FlyBase	http://flybase.org/ (v.FB2018_04, August 23, 2018)	McQuilton et al. (2012) St. Pierre et al. (2014) FlyBase Consortium (2003)
	FlyMine	http://www.flymine.org/ (version 46.0, 2018 July)	Lyne et al. (2007)
Population data	DGN	http://www.johnpool.net/genomes.html (v.1.1, July, 2016)	Lack et al. (2015) Lack et al. (2016)
	DGRP	http://dgrp2.gnets.ncsu.edu/ (v.2, February 2013)	Mackay et al. (2012) Huang et al. (2014)
	DPGP	http://www.dpgp.org/ (version 3, 2015)	Pool et al. (2012) Lack et al. (2015)
Interaction and pathway	DroID	http://droidb.org/ (v.2018_08, August 29, 2018)	Yu et al. (2008)

Table modified from Hales et al. (2015).

1.5. Objectives

This thesis is an integrative population genomics and evolutionary biology project following a bioinformatic approach that is performed in three sequential steps: (i) the comparison of five MKT approaches using empirical data from a North American population of *D. melanogaster* and simulated data, (ii) the inference of the evolutionary genome features influencing the evolution of protein-coding genes, (iii) the integration of patterns of genomic variation with annotations of large sets of spatio-temporal developmental data (evo-dev-omics) (Table 1.3).

Table 1.3 The three objectives of the thesis.

Objective	Outcome	Publication
Comparison of five MKT methodologies using both real and simulated data	Flowchart to select a MKT approach	Coronado-Zamora et al. (in preparation)
	iMKT: web server performing diverse MK-derived tests in *D. melanogaster* and human populations (`https://imkt.uab.cat`)	
	PopHumanScan: the online catalog of human genome adaptation (`https://pophumanscan.uab.cat`)	Murga-Moreno et al. (2018)
Inference of the evolutionary genome features influencing the evolution of protein-coding genes	Adaptive evolution is substantially impeded by Hill-Robertson interference in *Drosophila*	Castellano et al. (2015)
	Genomic features that increase and decrease the evolutionary capacity of proteins	Coronado-Zamora et al. (submitted, first result)
Integration of patterns of genomic variation with annotations of large sets of spatio-temporal developmental data	Adaptation and conservation throughout the *D. melanogaster* life cycle	Coronado-Zamora et al. (submitted, second result)
	Mapping selection within *D. melanogaster* embryo's anatomy	Salvador-Martínez et al. 2018

Gray-colored publications belong to other dissertations of the BGD group.

Comparison of five MKT methodologies using empirical and simulated data

The first objective is to estimate the rate of adaptive evolution, α, using five different MKT derived methodologies, as a benchmark to com-

pare their performance under real and simulated data. For that, we use genome-wide DNA variation data from a North American population of *D. melanogaster* and simulated data from SLiM 2 evolutionary simulation framework and test their performance in different conditions and evolutionary scenarios.

As a result of the efforts to represent, understand and interpret this huge amount of population genomic data, valuable resources have been created in collaboration with other members of the Bioinformatics of Genomics Diversity (BGD) group. Two resources have been developed: (i) iMKT, the integrative McDonald and Kreitman test web server (freely available at `https://imkt.uab.cat`), that allows performing diverse MKT derived tests on *D. melanogaster* and human populations (Coronado-Zamora et al., in prep.) and (ii) PopHumanScan, an online catalog that compiles and annotates all candidate regions under selection in the human genome (freely available at `https://pophumanscan.uab.cat`; Murga-Moreno et al., 2018).

Inference of the evolutionary genome features influencing the evolution of protein-coding genes

The second objective is to infer the genomic features that influence the evolution of protein-coding genes and discover genomic variation patterns. An inventory of genomic features has been estimated throughout the genome of *D. melanogaster*. These genomic features span four different characteristics of a genome: gene architectural, gene expression, genomic context and gene phylogenetic features.

These features are correlated with the population genomic parameters estimated in the first objective. The main aim is to assess how the features contribute to the genome adaptation and constraint and the observed patterns of genome variation.

Integration of patterns of genomic variation with annotations of large sets of -omics data

The third objective is to integrate the patterns of genome variation with annotations of large sets of spatio-temporal developmental data, with two main objectives:

TEMPORAL DIMENSION. The first objective is to measure the pattern of adaptive and selective constraint over the whole life cycle of *D. melanogaster* by integrating population genomics data with the complete developmental transcriptome of *Drosophila*. More specifically, the objectives are: (i) study whether *D. melanogaster* development follows the hourglass model or the von Baer's law, (ii) study whether there are differences not just in conservation but also in the rates of adaptive substitutions between stages and (iii) study whether the results in (i) and (ii) can be accounted by specific genomic features.

SPATIO-TEMPORAL DIMENSION. The second main objective is to carry out a global selection-phenotype-genotype integration, more specifically, to draw an exhaustive map of the selection acting on the complete embryo development of *D. melanogaster*. The specific objectives are: (i) estimate and compare both adaptation and selective constraint through the body of *D. melanogaster* and (ii) integrate the genomic features with the selection patterns during the embryo development.

Chapter 2

METHODOLOGY

2 Methodology

2.1. Data

2.1.1. Population genomics data

The present thesis has been carried out using *D. melanogaster* intraspecific variation data. The genomic sequences come from inbred isolines of a North American *D. melanogaster* population sequenced in the *Drosophila* Genetic Reference Panel (DGRP) project (Mackay et al., 2012). The processing and filtering of the genomic alignments and the estimation of the population statistics have been conducted by members of the Bioinformatics of Genomics Diversity (BGD) group, as participants of the DGRP consortium (Mackay et al., 2012; Huang et al., 2014). As a result of the efforts to represent, understand and interpret this huge amount of genomic data, valuable resources have been created, such as PopDrowser (Ràmia et al., 2012), the first *D. melanogaster* population genomics-oriented genome browser, which was the predecessor of PopFly (`https://popfly.uab.cat`), the most complete *D. melanogaster* population genomic browser up-to-date (Hervas et al., 2017) or iMKT (`https://imkt.uab.cat`), a web server that allows performing a battery of McDonald and Kreitman derived approaches (MKT) on *D. melanogaster* populations (Coronado-Zamora et al., in prep.).

Drosophila Genetic Reference Panel (DGRP)

The *D. melanogaster* population genomic data comes from the DGRP community resource (Mackay et al., 2012), a population consisting of 205 inbred lines originating from Raleigh, North Carolina, USA (Figure 2.1A). Two working phases (freezes) have been released of the DGRP, named "Freeze 1" and "Freeze 2". Freeze 1 was released in 2012 and the popu-

lation genomic analyses of its 168 sequences were published in Mackay et al. (2012). Freeze 2 was released in 2014 (Huang et al., 2014), enlarging the resource to 205 isolines.

In detail, Freeze 1 contains 168 isolines sequenced with a high coverage (average 22×, for details, see Mackay et al., 2012). 129 isolines were sequenced using Illumina technology (21.4× coverage), 10 using the 454 sequencing platform (12.1× coverage) and 29 using both technologies. The original 168 sequences were reduced to 158 because different contamination and duplications were detected in 10 isolines, and are not considered in the present thesis. Freeze 2 sequenced 48 DGRP lines that either were not previously sequenced or that only had 454 sequences available, and resequenced 6 isolines of the Freeze 1 that had a low coverage (Figure 2.1B). The average coverage reached was 27×. Using the population variability together with the genomes of two close species (*D. simulans* and *D. yakuba*), a battery of population statistics have been estimated in both releases which are the empirical starting point of all the work presented here.

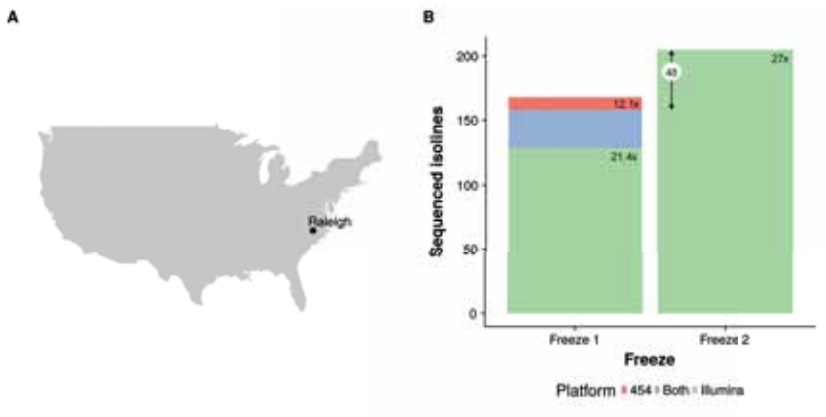


Figure 2.1 Population origin and overview of the sequenced isolines in Freezes 1 and 2. A. The DGRP population comes from Raleigh, North Carolina. **B**. 168 isolines were included in Freeze 1, 129 sequenced with Illumina (21.4× coverage), 10 with the 454 sequencing platform (12.1× coverage) and 29 using both platforms. Freeze 2 enlarged the resource to 205 isolines by sequencing 48 new lines and resequencing 6 more with an overall average of 27×.

The DGRP population was created collecting gravid females from a single population of Raleigh, followed by the full-sibling inbreeding approach during 20 generations to obtain full homozygous individuals (Figure 2.2). After this, the residual heterozygosis in the samples is ex-

pected to be 1.4% (inbreeding coefficient $F = 0.986$, Falconer and Mackay 1996). However, the expected 1.4% of residual heterozygosis was true for 90% of the sequenced chromosome lines. Huang et al. (2014) later found that 8% of the DGRP lines showed high values of residual heterozygosity (>9%) associated with large polymorphic inversions and they were excluded from the analyses.

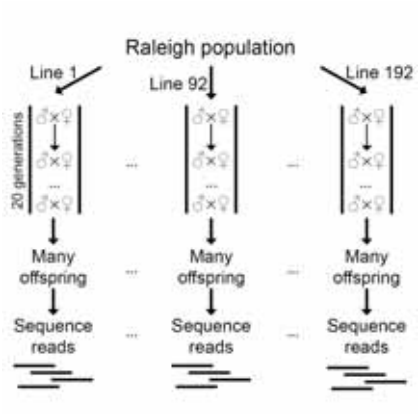


Figure 2.2 Experimental protocol for the creation and sequencing of the DGRP isolines. The DGRP population was created collecting gravid females around a farmer's market in Raleigh, followed by the full-sibling inbreeding approach during 20 generations to obtain full homozygous individuals. For each line, 500–1000 flies were extracted to perform high-throughput sequencing. Figure from Stone (2012).

The population genomic analyses performed using the Freeze 1 data used *D. yakuba* as outgroup species for the computation of divergence metrics. Furthermore, this dataset was reduced to 128 isolines due to the computational requirements of the method used to estimate selection (described in section 2.2.1). The analyses performed using the complete DGRP dataset, the Freeze 2 data consisting of 205 isolines, used *D. simulans* as an outgroup species and *D. yakuba* in some complementary analyses.

2.1.2. Gene expression data

The developmental expression data used in this thesis comes from two major resources: (i) the modENCODE consortium, from which the developmental transcriptome across *D. melanogaster* life cycle stages was obtained and (ii) Tomancak's embryogenesis expression dataset (Toman-

cak et al., 2007), a high-throughput database of mRNA expression in six different embryonic stages of *D. melanogaster*.

modENCODE data: temporal dimension

Gene expression data of 17,875 genes comes from RNA-seq experiments in the modENCODE project (Graveley et al., 2011) downloaded from FlyBase (release 6.06; last accessed: December 2015). In detail, the dataset contains the expression data for 30 stages of the whole life cycle of *D. melanogaster*, including 12 embryonic samples collected at 2-hour intervals for 24 hours, six larval, six pupal and three sexed adult stages at 1, 5 and 30 days after eclosion (Graveley et al., 2011). Reads per kilobase per million mapped reads (RPKM) values are provided only for exonic regions of the gene (excluding segments that overlap with other genes), except for genes derived from dicistronic/polycistronic transcripts, where all exon regions were used for the estimation of RPKM expression (see Gelbart and Emmert, 2013 for methodological details on the expression data processing).

BDGP data: spatio-temporal dimension

The Berkeley Drosophila Genome Project (BDGP) database was used to obtain the patterns of gene expression over the fly embryo's anatomy (Tomancak et al., 2007). This is a high-throughput database of mRNA expression spanning different embryonic stages. The BDGP divides the first 16 stages of embryogenesis defined by Hartenstein (1993) into six stages ranges: stage 1–3, stage 4–6, stage 7–8, stage 9–10, stage 11–12 and stage 13–16. This database has been the subject of previous studies dealing with computational image analysis of patterns of gene expression (Frise, Hammonds, and Celniker, 2010; Kumar et al., 2011; Salvador-Martínez and Salazar-Ciudad, 2015), but has never been combined with populational genomic data as in the present work. Based on the expert analysis of whole-mount *in situ* RNA-hybridization images, the BDGP database contains for each gene, the list of the embryonic anatomical structures in which it is expressed (`http://insitu.fruitfly.org/insitu/html/downloads.html/`; last accessed: December 2015). In detail, the BDGP has produced a large number of gene expression patterns, and textually annotated them with anatomical and developmental terms using a *controlled vocabulary* (cv). The terms spatially correspond to lo-

cal regions of the embryo and describe developmental and anatomical properties of gene expression.

2.2. Data analysis

2.2.1. Detecting and quantifying natural selection on gene coding regions

Inferring the action of natural selection on coding sequences relies on polymorphism and divergence data on two types of sites in the genome, one putatively selected and one neutral. Normally, non-synonymous positions (0-fold degenerated) are used as a proxy for selected sites, while synonymous positions (4-fold degenerated) are used as a proxy for neutral sites. This implies knowing the functional class of each nucleotide in the genome, which is not trivial. The same nucleotide can act, for example, as a coding site for a transcript, while as a UTR site for another. In this study, two criteria for the classification of each genome position were used: (i) a hierarchical criterion, used in the Freeze 1 data and (ii) the longest isoform criterion, used in the Freeze 2 data. With a hierarchical criterion, each position of the sequence is annotated following this order: 0-fold degenerated, 2-fold degenerated, UTR, intron, intergenic site and 4-fold degenerated (Figure 2.3A). With the longest isoform criterion, the longest CDS of a complete isoform is extracted and 0-fold and 4-fold degenerated sites are annotated accordingly. Therefore, UTR, intron, intergenic site and 2-fold degenerated sites are not considered (Figure 2.3B). In both cases, only synonymous and non-synonymous sites that are ortholog with the outgroup species were considered.

In the next paragraphs, more details about how the data was filtered and processed are given, however, a more complete description can be found in the original works published by members of the BGD group: Castellano et al. (2015) and Castellano (2016) filtered and processed Freeze 1 data (Mackay et al., 2012), and Hervas et al. (2017) and Coronado-Zamora et al. (in prep.) filtered and processed Freeze 2 data (Huang et al., 2014).

FREEZE 1 DATA. Coding exon annotations from *D. melanogaster* were retrieved from FlyBase (release 5.50; www.flybase.org; last accessed: March 2013). The number of synonymous (m_S), non-synonymous (m_N)

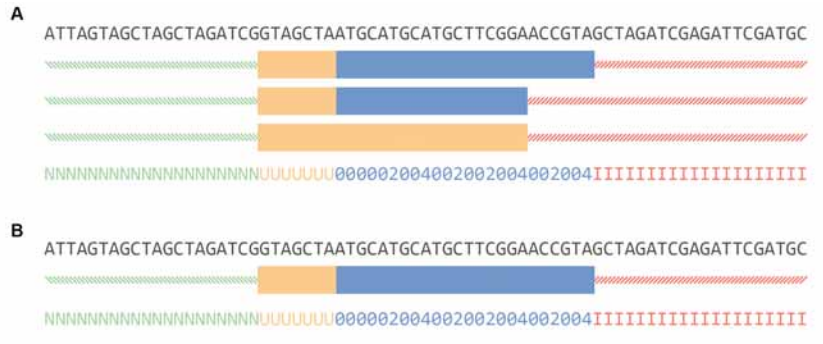


Figure 2.3 Examples of a recoded sequence. A. Recoding a sequence with multiple transcripts following a hierarchical criteria. **B**. Recoding the same sequence as A) but followiing the long isoform criteria. N is for intergenic (green); U is for UTR (orange); 0, 2 and 4 are the degeneracies of the coding regions (blue) and I is for introns (red).

and short introns (m_{ins}) sites and substitutions (P_S, P_N, P_{ins}, respectively) were computed. The folded site frequency spectrum (SFS) was calculated from the minor allele frequency (MAF, Figure 2.4A and 2.4B) for synonymous (SFS_S), non-synonymous (SFS_N) and short introns changes (SFS_{ins}) using an *ad hoc* Perl script. Divergence statistics (D) for synonymous (D_S), non-synonymous (D_N) and short intron sites (D_{ins}) were estimated using a multiple genomic alignment between DGRP isogenic lines (Mackay et al., 2012) and *D. yakuba* as outgroup species (Clark et al., 2007) using BDGP 5 coordinates (Berkeley *Drosophila* Genome Project 5; `www.fruitfly.org/sequence/release5genomic.shtml`), publicly available at `http://popdrowser.uab.cat` (Ràmia et al., 2012; last accessed: May 2010). Multiple hits were corrected using Jukes and Cantor correction (Jukes and Cantor, 1969).

One of the programs used for estimating adaptation (DFE-alpha, Eyre-Walker and Keightley, 2009) needs all sites sampled in the same number of sequences for all analyzed sites (Eyre-Walker and Keightley, 2009). Therefore, the original dataset of 158 lines was reduced to 128 by randomly sampling the polymorphisms at each site without replacement to accomplish the requirement. Finally, residual heterozygous sites and ambiguous positions (N) were excluded from the analysis. After the filtering, the dataset contained 11,103 protein-coding genes. In addition to using 4-fold degenerated sites as a proxy for the neutral mutation rate, short introns sites ($\leqslant$65 bp) were also used as an alternative neutral class. Following Halligan and Keightley's (2006) approximation, the positions

8–30 of introns shorter than 65 bp were used as a neutral reference. This final dataset consisted of 6,690 protein-coding genes.

FREEZE 2 DATA. For analyzing the sequences present in the complete DGRP release, the genome reference sequence and annotations corresponding to the FlyBase 5.57 release (http://flybase.org/; last accessed: September 2017) were used to assess the functional class of each genomic position. The *D. simulans* genome sequence and annotations were retrieved from FlyBase (release 2.0; Hu et al., 2013; last accessed: September 2017) and were used to estimate the derived allele frequencies (DAF or unfolded SFS, Figure 2.4A and 2.4C) and divergence metrics (D). Additionally, an alignment with *D. yakuba* (FlyBase release 1.3; Clark et al., 2007; last accessed: September 2017) was also used to compute the divergence metrics. The total number of fixed differences (D_N,

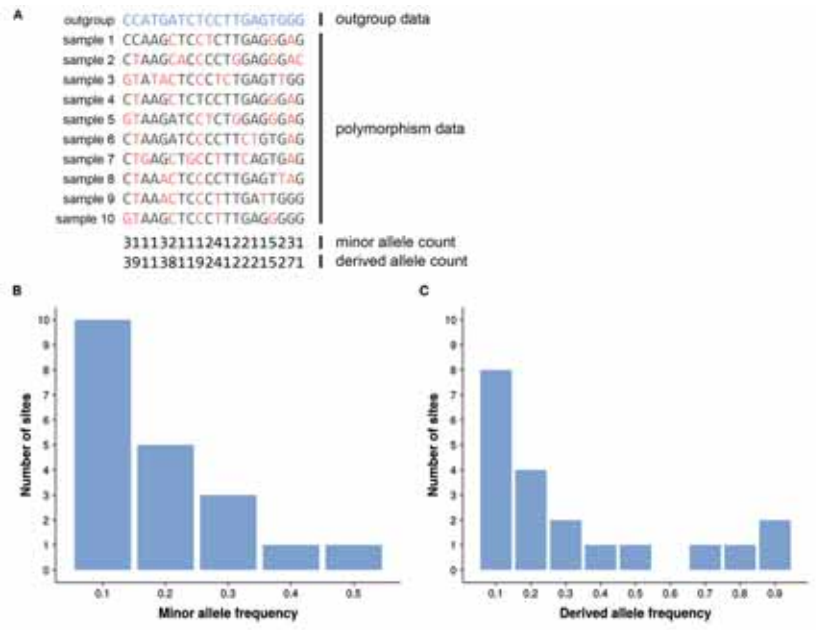


Figure 2.4 The site frequency spectrum (SFS). The number of polymorphic sites segregating at different frequencies in a population can be represented with the site frequency spectrum (SFS). **A**. 10 individual samples of a population and the corresponding outgroup sequence are shown. The numbers on the bottom part represent the counts for the minor allele and the derived allele. **B**. The minor allele frequency is used to obtain the folded SFS (frequencies ranging from 0 to 0.5). **C**. Additionally, the outgroup species sequence allows to know whether alleles are ancestral or derived, and the unfolded SFS or derived allele frequency can be obtained (frequencies ranging from 0 to 1). Figure adapted from Booker, Jackson and Keightley (2017).

D_S), polymorphic sites (P_N, P_S) and analyzed sites (m_N, m_S) in each site type were computed using *ad hoc* Perl scripts. Finally, polymorphism was categorized according to their frequency in 10 and 20 equally distributed bins to obtain the DAF. This dataset contains a total of 13,753 protein-coding genes.

Both Freeze 1 and Freeze 2 data represent a valuable population genomic resource which is the mainstay for the next analyses, the detection of the action of natural selection in protein-coding genes at the molecular level. Several methods for estimating adaptation and selection constraint that rely on polymorphism and divergence data have been applied, which are described in the next sections.

Standard McDonald and Kreitman test (MKT)

The standard McDonald and Kreitman test (MKT) is used to detect recurrent positive selection at the molecular level in a background of neutral mutations (McDonald and Kreitman, 1991). The standard MKT compares the amount of variation within a species (polymorphism, P) to the divergence between species (D) at two types of sites, one of which is assumed to evolve neutrally and is used as the null model to detect selection at the other type of site. In the standard MKT, these sites are synonymous (neutral, s) and non-synonymous sites (putatively selected, n) in a coding region. Under strict neutrality, the ratio of the number of selected and neutral polymorphic sites (P_N/P_S) is expected to be equal to the ratio of the number of selected and neutral divergence sites (D_N/D_S). The null hypothesis of neutrality is rejected when $D_N/D_S \neq P_N/P_S$. The excess of divergence relative to polymorphism for class n ($D_N/D_S > P_N/P_S$), is interpreted as a signature of adaptive selection on non-synonymous sites. Subsequently, the fraction of adaptive fixations (α) is estimated according to Equation 2.1. The significance of the test can be assessed with a Fisher's exact test on the 2×2 MKT contingency table (Table 2.1).

$$\alpha_{standard} = 1 - \frac{D_S}{D_N}\frac{P_N}{P_S} \tag{2.1}$$

Table 2.1 Standard MKT table.

Site class	Polymorphism	Divergence
Neutral	P_S	D_S
Selected	P_N	D_N

Fay, Wyckoff and Wu correction method (FWW method)

In the standard McDonald and Kreitman test, the estimate of adaptive evolution (α) can be easily biased by the segregation of slightly deleterious non-synonymous substitutions. Specifically, slightly deleterious mutations contribute more to polymorphism than to divergence, and thus, lead to an underestimation of α. Because they tend to segregate at lower frequencies than do neutral mutations, they can be partially controlled by removing low-frequency polymorphisms from the analysis. This approach is known as the Faw, Wyckoff and Wu method, FWW (Fay, Wyckoff, and Wu, 2001). In this case, α is estimated using the standard MKT equation, but considering only those polymorphic sites (for both neutral and selected classes) with a frequency above the established cutoff, typically 5%.

$$\alpha_{FWW} = 1 - \frac{D_S}{D_N}\frac{P_{N>5\%}}{P_{S>5\%}} \tag{2.2}$$

Extended MKT (eMKT)

An alternative approach that considers the presence of non-synonymous slightly deleterious mutations is the DGRP methodology (Mackay et al., 2012) and named here as extended MKT (eMKT). Because adaptive mutations and weakly deleterious selection act in opposite directions in the MKT, α and the fraction of substitutions that are slightly deleterious (b) will be both underestimated when both selection regimes operate (see Introduction, section 1.1.4, *MKT-based extensions*). To take adaptive and slightly deleterious mutations mutually into account, P_N, the count of segregating sites in class n, should be separated into the number of neutral variants and the number of weakly deleterious variants, $P_N = P_{N\,neutral} + P_{N\,weakly\,del.}$.

Table 2.2 eMKT table.

Site class	Polymorphism	Divergence
Neutral	P_S	D_S
Selected	$P_{N\,neutral}$	D_N

The estimate of the fraction of sites segregating neutrally within the DAF<5% ($f_{Neutral\,DAF<5\%}$) is $\widehat{f}_{Neutral\,DAF<5\%} = P_{S\,DAF<5\%}/P_S$. The expected number of segregating sites in the non-synonymous class n which are neutral within the DAF<5% is $\widehat{P}_{N\,neutral\,DAF<5\%} = P_N \times \widehat{f}_{Neutral\,DAF<5\%}$. The expected number of neutral segregating sites in the non-synonymous class n is $\widehat{P}_{N\,neutral} = \widehat{P}_{N\,neutral\,DAF<5\%} + P_{N\,DAF>5\%}$. Then, α is estimated substituting P_N with the expected number of neutral segregating sites, $\widehat{P}_{N\,neutral}$ (Table 2.2). The new equation is:

$$\alpha_{extended} = 1 - \frac{D_S}{D_N}\frac{\widehat{P}_{N\,neutral}}{P_S} \tag{2.3}$$

One advantage of the eMKT is the ability to quantify negative selection. For that, the excess of sites segregating at $DAF < 5\%$ with respect to the neutral site class is considered to be due to weakly deleterious segregating sites.

Asymptotic MKT

Messer and Petrov (2013) proposed a simple asymptotic extension of the MKT that yields accurate estimates of α, as it considers the presence of slightly deleterious mutations all along the DAF spectrum. Briefly, the asymptotic MKT first estimates α for each DAF category using its specific P_N and P_S counts and then fits an exponential (or a linear) function to these values, of the form: $\alpha_{fit}(x) = a + b^{(-cx)}$. Finally, the asymptotic α estimate is obtained by extrapolating the value of this function to 1:

$$\alpha_{asymptotic} = 1 - \frac{D_S}{D_N}\frac{P_{N(x)}}{P_{S(x)}} \tag{2.4}$$

$$\alpha_{asymptotic} = \alpha_{fit}(x = 1) \tag{2.5}$$

The BGD group proposed an extension of the asymptotic MK that estimates the fraction of deleterious substitutions following the extended MKT methodology, and refer to it as the integrative MKT (iMKT, Figure 2.5).

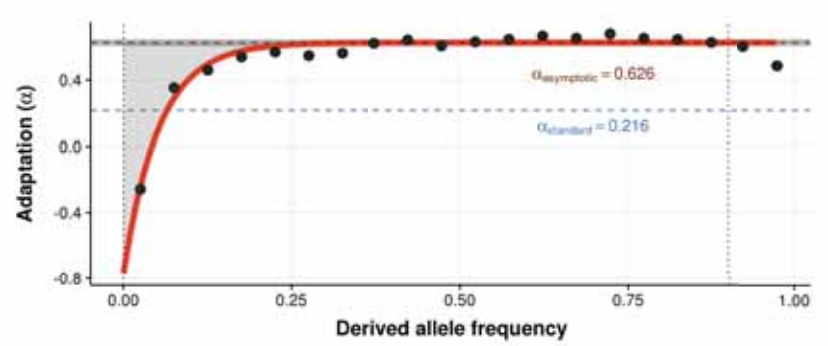


Figure 2.5 Example of results from iMKT using *D. melanogaster* 2R chromosome (Freeze 2) and *D. simulans* as outgroup. The two vertical lines show the limits of the *x* cutoff interval used (in the example [0,0.9]). Black dots indicate the binned α values for each DAF category following equation 2.4. The solid red curve shows the fitted $\alpha_{\text{fit(x)}}$. The dashed red line shows the final $\alpha_{\text{asymptotic}}$, following equation 2.5. The dark gray band indicates the 95% CI around the α estimation. The blue dashed line shows the estimated α using the standard MKT following equation 2.1 for comparison.

Simulations for testing the performance of MKT approaches

SLiM 2, a forward population genetic simulation software developed by Haller and Messer (2017), was used to test the different introduced MKT-derived methods. For this purpose, the SLiM configuration script provided on Messer's asymptoticMK's GitHub repository was used (available at `https://github.com/MesserLab/asymptoticMK`; last accessed: December 2017). The evolution of a population of 1,000 diploid individuals under 13 different scenarios was simulated, with 50 replications for each scenario to compute the simulated standard deviation ($\pm$SD). Simulation runs depended upon seven parameters: T, number of generations; L, chromosome length; μ, mutation rate; recombination rate; r_b, beneficial mutation rate; s_d, selection coefficient of deleterious mutation and s_b, selection coefficient of beneficial mutation. After an initial period of 10,000 generations to arrive at a steady-state (burn-in), runs executed for T additional generations. The simulated

chromosome was 10^7 bp long (L), with a uniform nucleotide mutation rate (μ) of 10^{-9} and an uniform recombination rate of 10^{-7} per base per generation. There are three different types of mutations: neutral type "$m1$" with a relative proportion of 0.5 of all new mutations and a selection coefficient (s) of 0; functional non-beneficial type "$m2$", with a relative proportion of 0.5 of all new mutations and a selection coefficient drawn from a gamma distribution with a mean s_d of -0.02 and a shape parameter of 0.2; and a functional beneficial type "$m3$", with a relative proportion r_b of 0.0005 and selection coefficient s_b of 0.1. The parameters described below (T, μ, r_b, s_b and s_d) were modified to run the 13 simulation scenarios proposed by Haller and Messer (2017). Fitness effects were assumed to be additive. Every 500 generations after the burn-in period, all polymorphisms were recorded in the population by dividing them according to their frequency into 20 equally sized frequency bins, and then adding them to an ongoing binned tabulation. At the end of each model run, binned values for the non-synonymous polymorphism P_N and synonymous polymorphism P_S were obtained. P_N was estimated from the combined mutations of types $m2$ and $m3$. P_S was estimated from all polymorphisms from mutations of type $m1$. Values for D_N and D_S were obtained from the set of mutations fixed during the simulation; i.e., D_N was estimated from the combined mutations of types $m2$ and $m3$ and D_S from all mutations of type $m1$. The output of SLiM 2 was used as input data for the different MKT methods to estimate α. The true value of α was estimated from the simulation run as the fraction of $d3/(d2+d3)$, where $d2$ is the number of $m2$ mutations fixed and $d3$ is the number of $m3$ mutations fixed. For this analysis, an x cutoff of [0.1,0.9] was used for estimating α with all the methodologies according to Haller and Messer (2017).

DFE-alpha

DFE-alpha (Eyre-Walker and Keightley, 2009), one of the most popular DFE-based methods, also corrects for the segregation of slightly deleterious alleles, providing a more accurate estimation than the standard MKT and other methods that do not take polymorphism data into account. Additionally, this method attempts to correct for possible effects of demography. For that, DFE-alpha incorporates two demographic situations: (i) a constant population size and (ii) a single, instantaneous change in population size from an ancestral size $N1$ to a present-day size $N2$ that occurred t generations ago. This software uses

a maximum-likelihood (ML) method based on polymorphism data to infer the distribution of fitness effects of new mutations. Like the MKT-based methods previously explained, DFE-alpha assumes two classes of sites in the genome: neutral sites (synonymous) and selected sites (non-synonymous) and contrasts the site frequency spectrum (SFS) at these two classes. As a neutral reference, two types of sites were used: 4-fold degenerated sites and the positions 8–30 of short introns ($\leqslant$65 bp) following Halligan and Keightley's (2006) approximation. As selected sites, 0-fold degenerated sites were used. Provided the SFS at both neutral and selected sites together with divergence data, the DFE-alpha method allows to calculate the proportion of fixed substitutions that are adaptive (α_{DFE}, equation 2.6) and the rate of adaptive substitutions relative to the neutral rate (ω_α, estimated as $\alpha_{DFE} \times \omega$, Gossmann, Keightley, and Eyre-Walker, 2012).

$$\alpha_{DFE} = \frac{d_N - d_S \int_0^\infty 2N\mu(N,s)f(s|a,b)\mathrm{d}s}{d_N} \tag{2.6}$$

Furthermore, in the analysis another complementary statistic was included, ω_{na} (estimated as ω-ω_a), which represents the proportion of non-adaptive substitutions (slightly deleterious and neutral) relative to the neutral rate (Galtier, 2016). Thus, the classical ω ratio is decomposed into these two metrics: ω_a and ω_{na}, and differentiate whether high rates of ω are due to positive selection or a relaxation of selection.

Therefore, natural selection on coding regions was estimated under a two-epoch demographic model and using a folded SFS (MAF). The SFS was folded to avoid difficulties with misidentification of the ancestral state (Hernandez, Williamson, and Bustamante, 2007) and because it performs well for inferring deleterious DFE (Eyre-Walker and Keightley, 2007; Boyko et al., 2008; Tataru et al., 2017).

Additionally, in some analyses the proportion of effectively neutral mutations was estimated (N_{es}<1) using the program `prop_muts_in_s_ranges.c` that comes with the DFE-alpha software.

Gene resampling (bootstrapping)

The different statistics to estimate the selection regimes parameters (i.e., α, ω, ω_a and ω_{na}) do not follow a parametric sampling distribution. For

this reason, we simulate the sampling distribution with the bootstrap method. The confidence intervals (CI) were obtained by estimating the selection regime parameters for 100 bootstrap replicates by sampling genes with replacement within each sampled bin.

The resampling was performed using the `boot` R package (Canty and Ripley, 2017) of R (R Core Team, 2017). Resampling is especially useful for DFE-alpha and the asymptotic methods, as they require the concatenation of several genes to compute α. This is because most genes do not have enough segregating or divergent sites to compute an MKT.

2.2.2. Gene expression through the developmental and life cycle stages

Gene expression data of 17,875 genes comes from RNA-seq experiments in the modENCODE project (Graveley et al., 2011). More in detail, the dataset contains the expression data for 30 stages of the whole life cycle of *D. melanogaster*, including 12 embryonic samples collected at 2-hour intervals for 24 hours, six larval, six pupal and three sexed adult stages at 1, 5 and 30 days after eclosion (Graveley et al., 2011).

Methodological limitations of the RNA-seq method together with experimental noise can lead to low, but positive RPKM values, even for not expressed genes. To account for this problem and just consider genes that are expressed in a certain stage, five different filter criteria were applied:

A LOW STRINGENT CRITERION, in which a gene is considered expressed in a stage if its RPKM is larger than zero. This criterion is used as a standard in other works (Hebenstreit et al., 2011; Wagner, Kin, and Lynch, 2013; Guillén, Casillas, and Ruiz, 2018). See Table B.1 for the genes analyzed in each stage for this criterion.

A LOW STRINGENT CRITERION WITH 2-FOLD DIFFERENTIAL EXPRESSION, in which a gene is considered expressed in a stage if its RPKM is larger than zero in that stage and if, in addition, its maximal gene expression (over all the stages) is at least twice that of its minimal gene expression (also over all the stages). See Table B.2 for the genes analyzed in the case of females and Table B.3 for the genes analyzed in males.

A LOW STRINGENT CRITERION WITH A 4-FOLD DIFFERENTIAL EXPRESSION, that is as the 2-fold criterion but with a 4-fold differential expression criterion. See Table B.4 for the genes analyzed in females and Table B.5 for the genes analyzed in males. These two criteria were analyzed separately for females and males. This is because their expression has been measured differently in the last three stages of modENCODE.

A MEDIUM STRINGENT CRITERION, in which a gene is considered expressed if its RPKM is equal or higher than 2. See Table B.6 for the genes analyzed in each stage.

HIGH STRINGENT CRITERION, in which a gene is considered expressed if its RPKM is equal or higher than 10. This criterion is also used as a stringent criterion in RNA-seq analysis by other authors (Dezso et al., 2008). See Table B.7 for the genes analyzed in each stage.

Figure 2.6 shows the number of genes for each criterion and stage. As a result of applying each of these five criteria on the same modENCODE RNA-seq data, seven different lists of genes for each life cycle stages were obtained (Tables B.1–B.7).

Additionally, for all previous analyses, genes that were constitutively expressed in all stages were discarded. In an additional analysis, genes that are always expressed in all stages under the low stringent criterion were also considered (6,655 genes, from which 5,687 can be analyzed using 4-fold sites as a proxy for the mutation rate and 3,758 can be analyzed using short intron sites as a proxy for the mutation rate). Finally, individual genes having polymorphic and divergent sites were also analyzed using MKT derived methods (see Table B.8 for the genes considered).

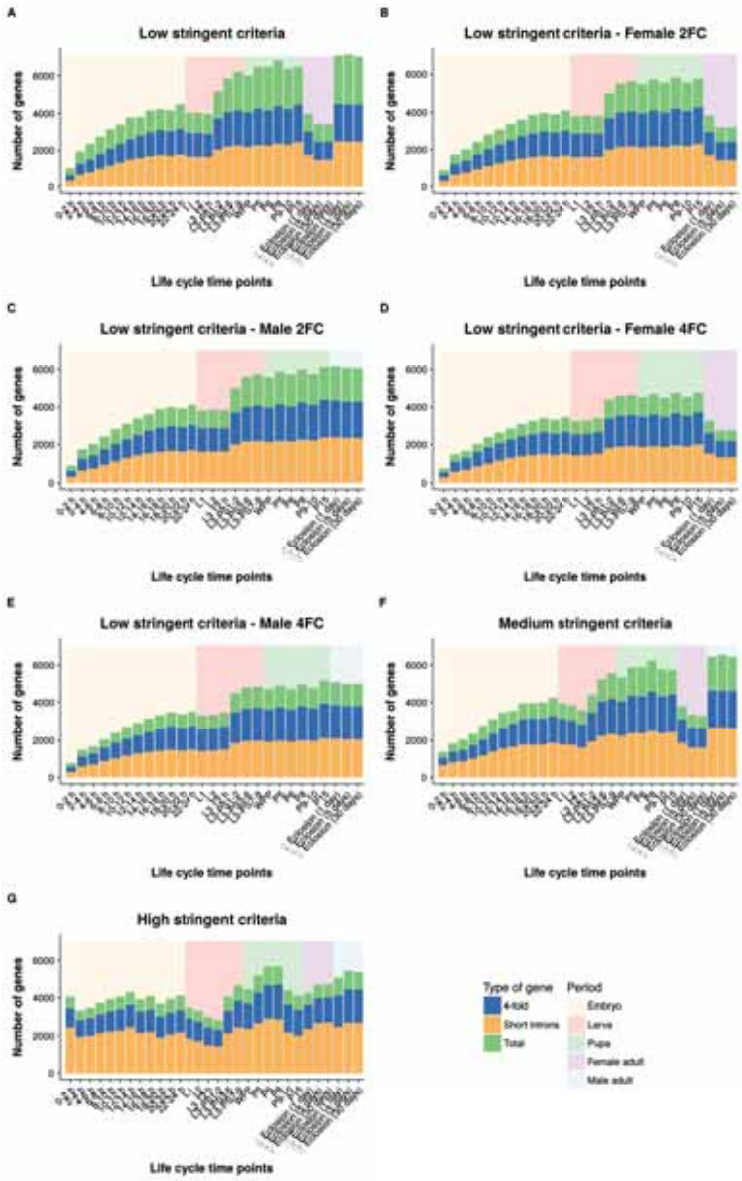


Figure 2.6 Genes expressed for each criterion and stage. The title of subfigures **A**-**G** denotes the criterion used. In green, total genes expressed in each stage. In blue, the proportion of genes that can be analyzed with 4-fold data. In yellow, the proportion that can be analyzed with short intron data. The embryonic stages are named by the time intervals (from 0h to 24h), the larval stages are the first instar (L1), second instar (L2) and third instar (L3). The L3 stages are subdivided into the first 12 hours (L3-12h) and several puff stages (L3-PS1 to L3-PS7). WPP is the white pre-pupae stage. The pupal stages with RNA-seq are phanerocephalic pupa, 15h (P5), 25.6 hours pupa (P6), yellow pharate, 50.4 hours (P8), amber eye-pharate, 74.6 hours (P9-10), green meconium pharate, 96 hours (P15). Adult stages are 1, 5 and 30 days after eclosion (1 day, 5 days and 30 days).

Gene expression profile clustering

To identify shared temporal expression patterns among the genes of the modENCODE RNA-seq experiments, a soft clustering method to the $\log_2$-transformed RPKM expression values was applied. Gene expression clusters are not well defined in expression time-course data, as is the case, and soft clustering methods are then advised to identify clusters (Futschik and Carlisle, 2005). A fuzzy c-means algorithm with the `mfuzz()` function of the R package `Mfuzz` (Futschik, 2015) was used. The Mfuzz soft clustering algorithm uses the Euclidean distance as a distance metric and requires two main parameters: c, the number of clusters and m, the fuzzification parameter. For the clustering, $\log_2$-transformed expression values were z-standardized, so that the average expression value for each gene is zero and the standard deviation is one. The fuzzy soft clustering method is different from hard clustering (like hierarchical clustering) in the sense that genes are not uniquely assigned to one cluster. Instead of this, a gene i has gradual degrees of membership μ_{ij} to a cluster j. High membership values indicate a high correlation between gene i with the cluster centroid c_j (Futschik, 2015). The `mfuzz()` function uses the fuzzy c-means algorithm, based on minimization of a weighted square error function with which the clusters centroids c_j result from the weighted sum of all cluster members. The membership value (μ_{ij}) indicates how well the gene i is represented by cluster j. Genes having a cluster membership lower than 0.8 were excluded. The c and m values were optimized using the procedure described in Futschik (2015) and Futschik and Carlisle (2005), resulting in $c = 9$ clusters for both datasets and a m parameter of 1.23 and 1.08 for the embryonic development (Figure 2.7) and life cycle (Figure A.1), respectively. Therefore, for the embryo development, this resulted in 9 different clusters based on the expression pattern of 3,819 embryo-expressed genes, out of 5,514 embryo-expressed genes determined with the low stringent criteria. One of the clusters was discarded as it consisted of very few genes with a membership value $\geqslant 0.8$ (90 genes), and therefore only 8 were finally analyzed (Figure 2.7). In the case of the genes expressed in the whole life cycle, 9 clusters were obtained based on the expression of 8,167 genes, out of 9,241 genes expressed in the whole life cycle (according to the low stringent criteria and discarding female expression data, so genes exclusively expressed in adult females were not considered). Tables 2.3 and B.9 show the number of genes expressed in each cluster for the two analyses.

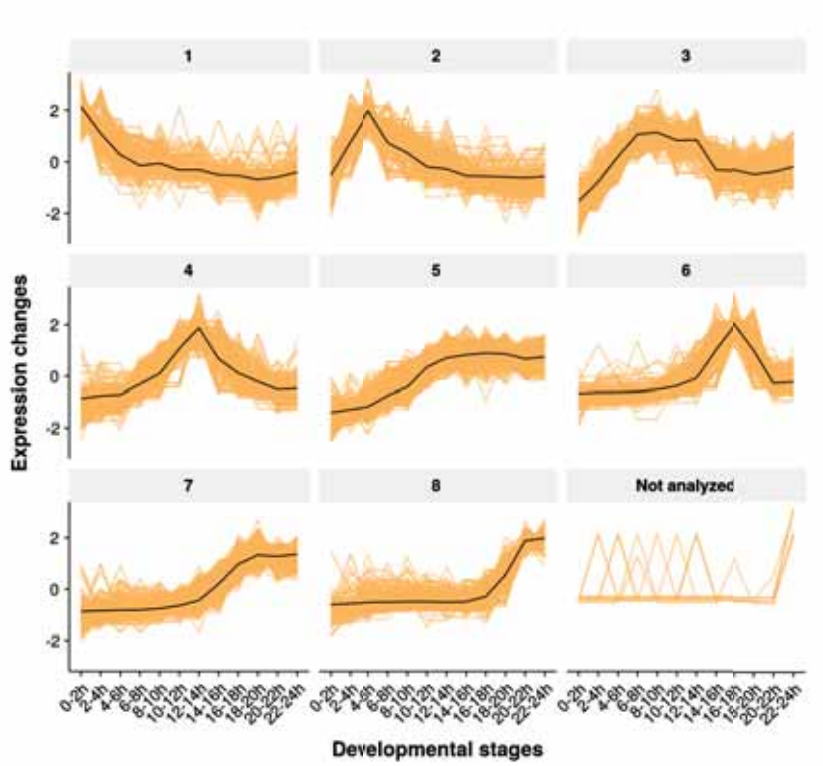


Figure 2.7 Temporal profile of expression of the genes in each embryonic development cluster. Each yellow line represents the expression pattern of a given gene. The black line represents the average expression of all genes expressed in a stage. All shown genes have a cluster membership $\geqslant 0.8$. Cluster 9 was not analyzed as it consisted of very few genes with a membership value $\geqslant 0.8$.

Maternal, maternal-zygotic and zygotic genes

A list of maternal, maternal-zygotic and zygotic genes was obtained from data by Thomsen et al. (2010) using egg and early development microarray analyses. *Maternal* genes are those genes which mRNA is shed in the egg and which are not transcribed by the embryo. *Maternal-zygotic* genes are those genes which mRNA is shed in the egg by the mother but that are also transcribed by the embryo. *Zygotic* genes are genes which mRNA is exclusively transcribed by the egg. The maternal gene list was obtained joining the original Thomsen's categories for non-transcribed genes: "maternal decay", "mixed decay", "stable" and "zygotic decay" categories (4,942 genes). The maternal-zygotic list was created by joining

Table 2.3 Genes expressed in 8 clusters of the embryo development. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Cluster	Total genes	Membership $\geqslant 0.8$	4-fold genes	Short-intron genes
1	311	229	140	79
2	540	403	198	99
3	614	303	142	68
4	519	315	201	121
5	845	597	500	324
6	411	272	184	95
7	1,325	1,096	837	453
8	785	604	419	226
9	-	-	-	-
Total	5,350	3,819	2,621	1,465

the categories of genes that are both present in the egg and transcribed later (1,332 genes analyzed): "maternal decay + transcription" and "stable transcription" categories. Finally, the zygotic genes correspond to the original "purely zygotic" category (850 genes). Three lists of genes were obtained, one for the maternal genes, one for the maternal-zygotic and one for the zygotic genes. See Table 2.4 for the genes analyzed in each category.

Table 2.4 Maternal, maternal-zygotic and zygotic genes analyzed. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Class	Total genes	4-fold genes	Short-intron genes
Maternal	4,942	4,255	2,808
Maternal-Zygotic	1,332	1,162	740
Zygotic	850	690	359
Total	6,992	5,999	3,836

For assessing if a developmental stage, cluster or gene category undergoes differential selection compared to the genes not expressed in such group of genes, a permutation test was applied. Table 2.7 summarizes all the hypotheses tested with this data and in section 2.3.1 the methodological details about the performed permutation test are explained.

2.2.3. Anatomical structure data

For each gene available in the BDGP database, the list of the embryonic anatomical structures in which it is expressed was obtained (BDGP; http://insitu.fruitfly.org/insitu/html/downloads.html/; last accessed: December 2015). Gene IDs were validated and updated using FlyBase converting id tool obtaining a total of 5,969 genes (http://flybase.org/static_pages/downloads/IDConv.html; last accessed: December 2016). The gene ID of the polymorphism and divergence datasets were also validated and updated, obtaining 11,074 genes for the 4-fold gene dataset (instead of 11,103) and 6,671 genes for the short-intron dataset (instead of 6,690).

The original anatomical structure dataset was collapsed into 18 different anatomical structures as described in Tomancak et al. (2007) and only genes without "no staining" as their unique term were analyzed. Some of the structures are visually displayed in Figure 2.8.

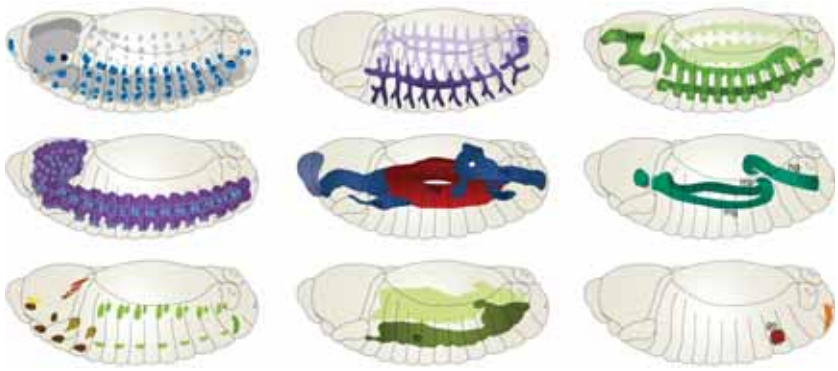


Figure 2.8 Visual representation of anatomical structures analyzed in this study. From left to right and from top to bottom: "Peripheral nervous system" (PNS), "Tracheal system", "Head mesoderm" (*hms*, head mesoderm), "Procephalic ectoderm/-Central nervous system (CNS)", Intestinal tract (including: "Foregut", "Salivary gland", *sg*, salivary glands; "Midgut", *mg*, midgut; "Hindgut", *hg*, hindgut), Visceral musculature (including "Malpighian tubules", *mp*, Malpighian tubules; *mg*, midgut; *hg*, hindgut), "Ectoderm/Epidermis", "Fat body", "Germ line" (*go*, gonads). Not shown, but analyzed: "SNS", "Optical lobe", "Segmental/GAP", "Garland cells/Plasmatocytes/Ring gland", "Yolk", "Circulatory system", "Ubiquitous", "Maternal". Images modified from Hartenstein (1993) with permission.

The anatomical structure classification was done collapsing all the data globally (thus, only considering the spatial dimension, Table 2.5) and dividing it into six embryo developmental stages (considering the spatio-

temporal dimension). The number of genes analyzed is shown in Table B.10.

Table 2.5 Genes expressed in each anatomical structure. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Antomical structure	Origin	Total genes	4-fold genes	Short-intron genes
Ectoderm/Epidermis	Ectoderm	1,521	1,294	726
Foregut	Ectoderm	1,062	912	502
Hindgut/Malpighian tubules	Ectoderm	1,171	1,023	608
PNS	Ectoderm	353	298	160
Procephalic ectoderm/ CNS	Ectoderm	1,731	1,490	910
Salivary gland	Ectoderm	320	285	174
SNS[#]	Ectoderm	54	47	20
Optical lobe[#]	Ectoderm	154	126	58
Tracheal system	Ectoderm	581	485	251
Endoderm/Midgut	Endoderm	1,892	1,654	1,024
Garland/Plasmatocytes/ Ring gland	Mesoderm	467	414	246
Head mesoderm/Circ. Syst./FB	Mesoderm	724	640	374
Mesoderm/Muscle	Mesoderm	1,449	1,270	767
Ubiquitous	Other	2,956	2,589	1,682
Maternal	Other	4,283	3,762	2,382
Germ line	Other	480	419	281
Segmental/GAP[#]	Other	149	130	68
Amnioserosa/Yolk	Other	477	419	228
Total		5,671	4,945	3,028

[#] Anatomical terms that were not analyzed in posterior analyses (not enough genes to be analyzed, the minimum is 150 genes).

Germ layer data

Genes were further classified by the germ layer they are derived from: ectoderm, endoderm and mesoderm (Figure 2.9). For example, the cv

"dorsal epidermis" is classified as ectoderm. A gene was assigned to a certain germ layer if it was expressed only in an anatomical structure belonging to it (and not in those anatomical structures of any other germ layer). Table 2.6 contains the genes expressed in each germ layer.

Table 2.6 Genes expressed in each layer. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Layer	Total genes	4-fold genes	Short-intron genes
Ectoderm	1,324	1,137	663
Endoerm	303	269	168
Mesoderm	302	271	168
Total	1,929	1,677	999

For assessing if an anatomical structure or gene layer undergoes differential selection compared to the genes not expressed in such anatomical structure or gene layer, a modification of the standard permutation test procedure was applied. The novel introduction of such a permutation test for this data is further discussed in section 4.3.3. Table 2.8 summarizes all the hypotheses tested with this spatio-temporal dataset and in section 2.3.2 methodological details about the performed permutation test are explained.

Figure 2.9 Visual representation of the embryonic germ layers. The mesoderm is represented in red, the ectoderm in blue and the endoderm in yellow. The amnioserosal covering of the embryo is represented in white. Figure taken from Gilbert, 2014.

2.2.4. Genomic features metrics

In this section, the different genomic features analyzed in the present thesis are defined and details about how they were measured are given. These measures span from gene architectural features to expression level and local recombination rate estimates. The analyses were performed using the Freeze 1 dataset, with FlyBase release 5.50 annotations.

A total of four architectonic features were measured: gene size, number of exons, number of transcripts and intron length. *Ad hoc* Perl and bash/awk scripts were implemented for the calculations.

GENE ARCHITECTURAL FEATURES. Features that are related to the gene structure, including *gene size*, the *number of exons and transcripts* and *intron length*.

Gene size. Defined as the length of the coding sequence (CDS) of a gene in bp.

Number of exons and transcripts. Number of different exons and transcripts of a gene, respectively. It is the total count of all exons and transcripts that are annotated for a gene in the FlyBase.

Intron length. It is the average distance between the exons of a gene in bp. If two exons overlap, the longest was used for the estimation. Each gene is treated as an independent entity, therefore, overlapping genes were not taken into account. Figure 2.10A represents a scheme of the estimation of the average intron length of a gene with three transcripts.

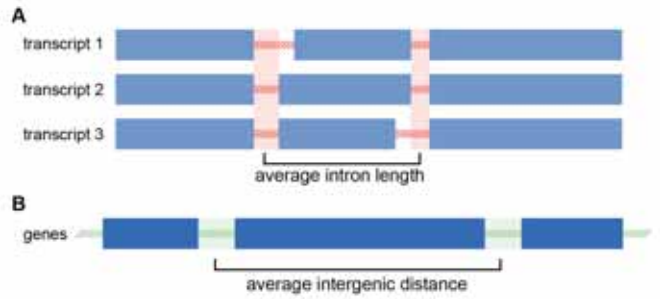


Figure 2.10 Example of intron and intergenic distance measurement. Red and green highlighted parts are used to estimate the intron length and intergenic distance, respectively. **A**. Light blue boxes represent exons. The conserved intronic regions of transcripts are used to estimate the average intron length. **B**. Dark blue boxes represent genes. For the estimation of the average intergenic distance, the average distance between the two closest genes from a given gene is computed.

GENE EXPRESSION FEATURES. Features that are related to the transcription profile of genes, which includes the expression bias, expression level, codon usage and the transcriptome divergence index.

Expression bias and expression level. Expression bias and expression level were estimated using the modENCODE data, described in section

2.2.2. Expression bias (or breadth) was estimated for this dataset using Yanai et al.'s (2005) tissue specificity metric, τ:

$$\tau = \frac{\sum_{j=1}^{n} 1 - \frac{\log(S_j)}{\log(S_{max})}}{n-1} \tag{2.7}$$

where S is the logarithm of the RPKM and n is the number of developmental stages. τ ranges from 0 to 1. A gene with a τ value close to 0 indicates that it is broadly expressed across multiple stages or tissues. A gene with a τ value close to 1 indicates that it has a highly biased expression. A τ equal to 1 means that a gene is expressed in only one stage or tissue.

Average gene expression level is measured as the logarithm of the average expression in the 30 stages in RPKM units. In some analyses, the maximum expression level was used, measured as the logarithm of the maximum expression in the 30 stages in RPKM units.

Codon bias. Measured as the frequency of optimal codons, *Fop*. The software CodonW was used for the estimation of *Fop* (Peden 1999; www.codonw.sourceforge.net; last accessed: June 2012). A specific *D. melanogaster* codon table already provided by this software was used. The index is estimated as the ratio of optimal codons to synonymous codons. Values range between 0 (no optimal codons are used) and 1 (only optimal codons are used).

Transcriptome divergence index. The transcriptome divergence index (TDI, Quint et al., 2012) measures the sequence divergence but weighted by the relative expression level of a gene. It is a measure of the average transcriptome selection pressure, estimated as:

$$TDI_s = \frac{\sum_{i=1}^{n} \frac{K_{ai}}{K_{si}} e_{is}}{\sum_{i=1}^{n} e_{is}} \tag{2.8}$$

where n is the total number of genes i analyzed in each stage s, and e is the expression level of the gene (as the logarithm of the RPKM).

GENOMIC CONTEXT FEATURES. Features that are related to the relative position of genes in a genome, including the *recombination rate*, the *intergenic distance* and the gene density.

Recombination rates. Recombination rate estimates at 100 kb non-overlapping windows were retrieved from Comeron, Ratnappan, and Bailin (2012). This data was used to assign exons to the recombination rate of the window they are located. Gene recombination rate consists of the average exons recombination rate. Briefly, these empirical recombination rate estimates consist of the characterization of the products of 5,860 female meioses by genotyping a total of 139 million informative SNPs and mapped 106,964 recombination events at a resolution up to 2 kilobases. These are the most precise and comprehensive estimates of recombination available for *D. melanogaster* up to date.

Intergenic distance. It is the average distance between the two closest genes to a given gene in bp. When a gene is nested in another one, the intergenic distance is set as 0. Figure 2.10B represents a scheme of the estimation of the average intergenic distance of a gene.

Gene density. Additionally, gene density was also used for some of the conducted analyses. The data was retrieved from the study in Castellano et al. (2015). To compute gene density the midpoint coordinate of each gene was calculated first. The start point corresponds to the first position of the first coding exon and the stop point corresponds to the last position of the last coding exon. Then all coding sites 50 kb upstream and 50 kb downstream the midpoint coordinate were counted and this coding sequence count was used as an estimate of gene density.

GENE PHYLOGENETIC FEATURES. Features related to the phylogeny of genes, which includes the phylogenetic age.

Phylogenetic age. A phylogenetic age was assigned to each gene using the phylostratigraphic maps of *D. melanogaster* from Drost (2014). These maps assign a phylogenetic age to each protein-coding gene in a species of interest (in this case *D. melanogaster*) based on the phylogenetic level at which orthologs for that gene are found (e.g., if a gene has orthologs at the level of eukaryota, the phylogenetic age is older than if a gene has only orthologs among Drosophilids). With this method, each gene can be assigned a discrete age category ranging from 1 to 13, or phylostratum (PS), corresponding to hierarchically ordered phylogenetic nodes along the tree of life database (Figure 2.11, Drost et al., 2015). The PS dataset was downloaded from `http://dx.doi.org/10.6084/m9.figshare.1244948/` (last accessed: May 2015). As this data set uses FlyBase protein IDs as identifiers, the R packages `biomaRt` (Durinck et al.,

2005) and `AnnotationDBI` (Pagès et al., 2017) were used to convert them into FlyBase Gene IDs.

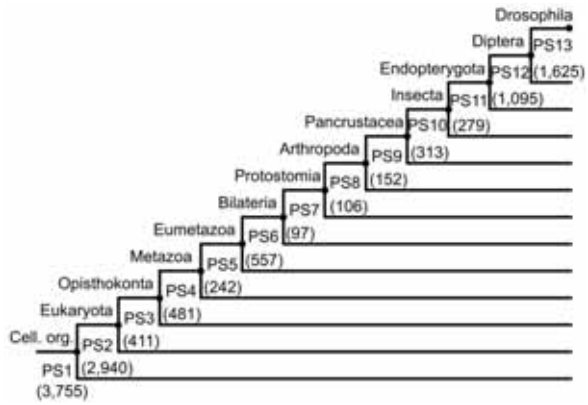


Figure 2.11 *D. melanogaster* phylostratigraphic map. Numbers in parenthesis represents the number of genes per phylostratum (PS1-PS13). Figure taken from Drost et al. (2015).

Table B.11 contains a summary statistic of the features analyzed and Figure 2.12 depicts the features distribution.

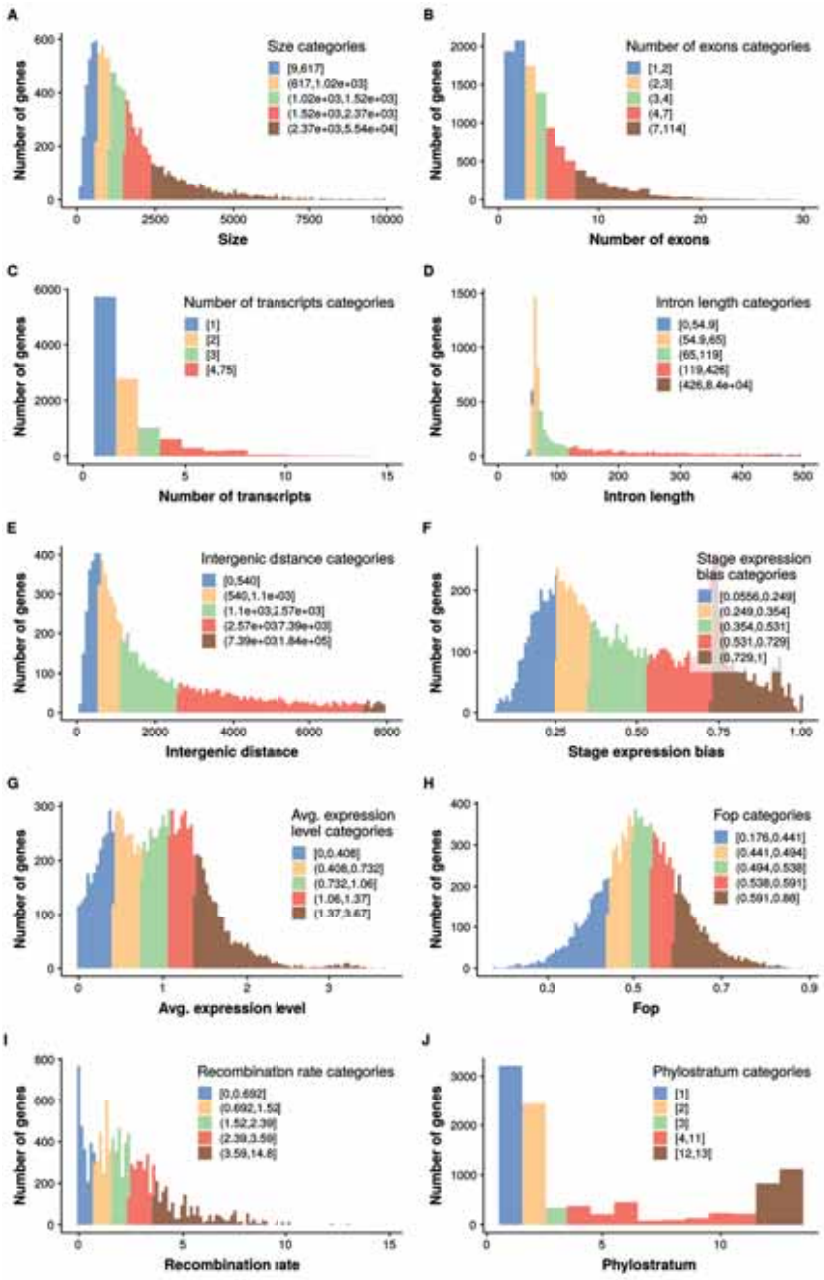


Figure 2.12 Features distribution for the *D. melanogaster* gene dataset. **A**. Size distribution and categories. **B**. Number of exons distribution and categories.**C**. Number of transcripts distribution and categories. **D**. Intron length distribution and categories. **E**. Intergenic distance distribution and categories. **F**. Expression bias distribution and categories. **G**. Expression level distribution and categories. **H**. *Fop* distribution and categories. **I**. Recombination rate distribution and categories. **J**. Phylogenetic age distribution and categories.

2.2.5. Testis and immune genes

The Gene Ontology (GO) terms were downloaded for the gene dataset through the R package `biomaRt` (Durinck et al., 2005) using the *D. melanogaster* ENSEMBL database, following the procedure used at Castellano et al. (2015). When a gene was associated to any term related to testis or the immune system (see Table B.12 for the related genes to this terms) it was removed from the expression dataset of the low stringent criterion (a total of 171 out of 2,869 genes were removed). Those 171 exhibit higher rates of adaptation as it would be expected (permutation test, ω_a, p-value = 0.028; ω, p-value < 0.001) when compared against the complete *Drosophila* dataset (6,690 short intron genes).

2.2.6. GO enrichment

Gene Ontology (GO) enrichment analysis was performed using the PANTHER Overrepresentation Test (Release 20171205, Mi et al., 2017) applying a Fisher's exact test with a false discovery rate (FDR) multiple test correction (p-value < 0.05).

2.2.7. Standard statistical analysis

Correlations between temporal profiles were carried out by Spearman's rank correlations, calculated by the `cor.test()` function in R (R Core Team, 2017).

Analyses of variance (ANOVA) were performed using the `lm()` and `anova()` functions of R (R Core Team, 2017). The homogeneity of variances was assessed with the Fligner-Killeen test, implemented in the `fligner.test()` function of R (R Core Team, 2017).

2.3. Statistical analysis

In the following section, details about the different statistical tests applied in this thesis are given. At the end of this section, two tables (Tables 2.7 and 2.8) summarize the hypotheses tested.

2.3.1. Permutation test for temporal analysis

To assess whether developmental stages or gene clusters undergo differential selection compared to the genes not expressed in such stage or gene cluster, a permutation test was applied. For obtaining a null distribution for the differences between gene groups, the complete list of genes was shuffled without replacement 1,000 times via *ad hoc* bash and Perl scripts. Adaptation rate and selective constraint were estimated in each randomized list, obtaining an expected null distribution. The two-tailed *p*-value was obtained by counting the number of replicates below and above the observed difference divided by the total number of replicates (i.e., 1,000). Multiple comparisons for each analysis were corrected by the false recovery rate (FDR) approach (Benjamini and Hochberg, 1995).

See Figure 2.13 for a graphical summary of the permutation procedure and Table 2.7 for a summary of the hypothesis tested in the second part of the thesis that used the classical permutation test procedure.

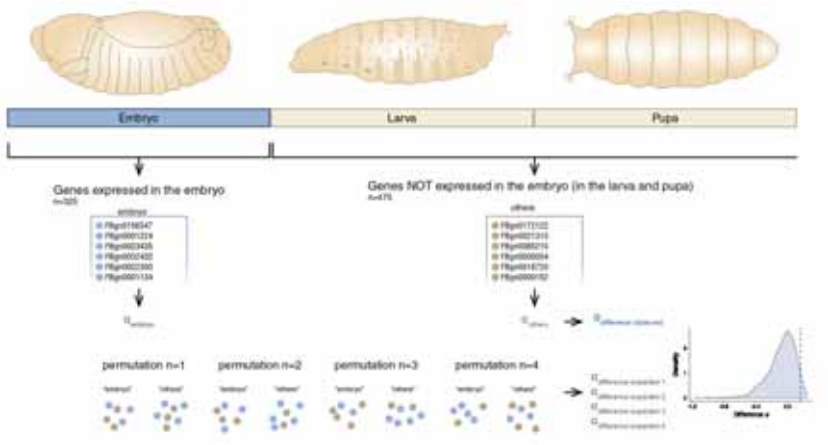


Figure 2.13 Classical permutation test procedure performed for the temporal analysis. First, the gene datasets should be defined. In the figure, these datasets are the sets of genes that are expressed in different developmental stages. The genes are divided into two datasets depending on the hypothesis to test. The hypothesis in the figure is whether genes that are expressed in the embryo development undergo differential expression compared to the genes that are not expressed in the embryo development. The statistics of interest are estimated for the two datasets. In the figure, the difference of the proportion of adaptive substitutions (α_{DFE}) is estimated. Next, the same statistic is calculated in each permuted dataset: genes are labeled as "embryo" and "others" randomly 1,000 times. Finally, the significance of the observed statistic is obtained by comparing it with the distribution of the expected statistic.

Table 2.7 Summary of the permutation test performed in the temporal study. The number of genes in the *Null distribution* column correspond to the ones analyzable with short-intron sites.

Analysis	Hypothesis tested	Null distribution	Comparisons
Developmental periods	For a given developmental period, are the genes expressed in one developmental period undergoing differential selection compared to the genes not expressed in such developmental period?	Genes expressed during the whole development (2,869 genes using the low stringent criterion) or genes in the complete dataset (6,690 genes).	8
Embryo development expression clusters	For a given embryo development cluster, are the genes expressed in one cluster undergoing differential selection compared to the genes not expressed in such cluster?	Genes expressed during the embryo development (2,012 genes using the low stringent criterion).	8
Life cycle expression clusters	For a given life cycle cluster, are the genes expressed in one cluster undergoing differential selection compared to the genes not expressed in such cluster?	Genes expressed during the whole development (E-L-P-Males) (female-biased genes where excluded) (2,860 genes using the low stringent criterio).	9
Maternal, maternal-zygotic, zygotic	Are the genes expressed in a category undergoing differential selection compared to the genes not expressed in that category?	Set of maternal, maternal-zygotic and zygotic genes (3,836).	3
Immune and testis genes	Are immune and testis-related genes undergoing adaptive selection compared to the other genes in the dataset?	Genes in the complete gene dataset (6,690 genes).	0

2.3.2. Permutation test for spatial analysis

To assess whether anatomical structures or germ layers undergo differential selection compared to other genes, a permutation test was applied. Specifically, a matrix was first built, in which each column represents an anatomical structure or layer and each row represents a gene. The matrix is filled with 0 and 1, with 0 indicating no expression and 1 indicating expression of each gene in an anatomical structure or germ layer. To generate the expected null distribution, the gene ID labels in the matrix are reshuffled at random. Each reshuffle of the labels represents a new permuted dataset in which genes are distributed randomly between anatomical structures (or germ layers) while keeping the number of genes per anatomical structure (or germ layer) constant. Therefore, in each permuted dataset, the number of genes co-expressed between each anatomical structure is as in the original dataset. This allows inferring the null distribution of the statistical output (α, ω, ω_a, ω_{na}) simultaneously for all the anatomical structures. This captures the correlational structure of the data and, contrary to the previous permutation test (section 2.3.1), there is no need to compute as many null distributions as hypothesis drawn. This results in fewer permutation tests to run and more statistical power. The reshuffling process was repeated 1,000 times to obtain the null distribution. A two-tailed p-value was obtained by counting the number of replicates above or below the observed value, dividing the value by the total number of replicates (1,000) and multiplying it by 2 (Figure 2.14).

The following Table 2.8 summarizes the hypothesis tested in the third part of the thesis that used the permutation test procedure for spatial analysis.

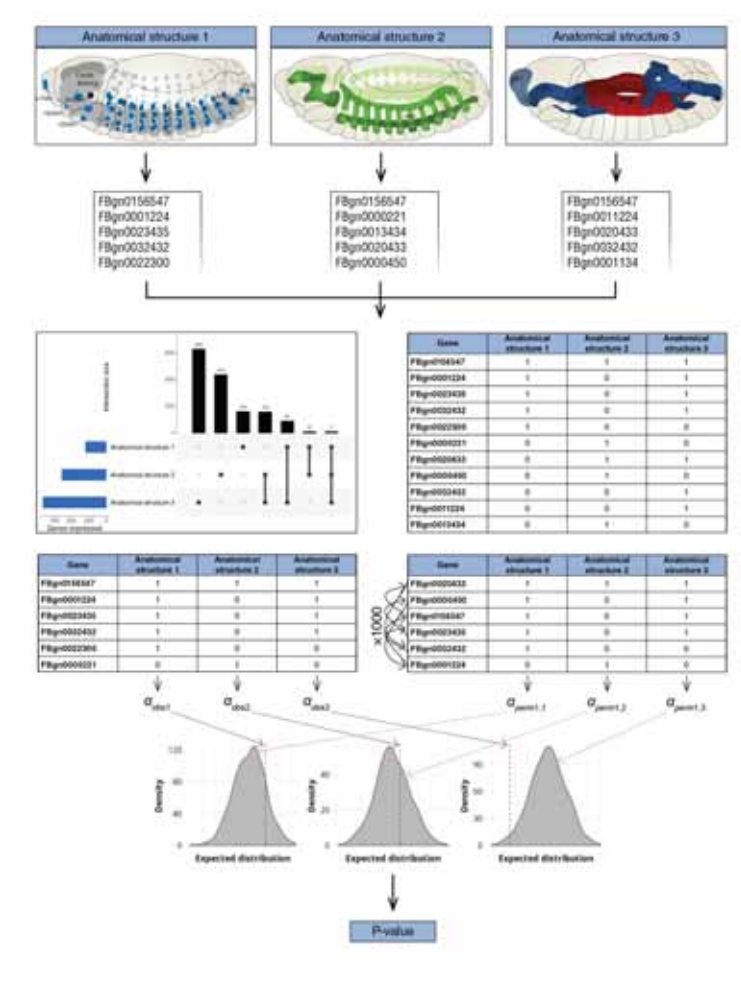


Figure 2.14 Permutation test procedure for the spatial analysis. First of all, the anatomical structures to study are defined. In the figure, the peripheral nervous system (PNS), head mesoderm and salivary glands are shown. The list of the genes expressed in each anatomical structure of study is obtained. A matrix in which each column stands for an anatomical structure and each row stands for a gene is built. The total number of rows is equivalent to the number of genes in the dataset. The matrix is filled with 0 and 1, with 0 indicating no expression and 1 indicating expression. This represents the genes shared between the anatomical structures. Once the matrix is built, the observed values of an statistic are calculated (an estimation for each anatomical structure *i*, $\alpha_{1..i}$). To generate the expected distribution, the gene ID labels in the matrix are randomly reshuffled. Each reshuffle *j* of the labels represents a new permuted dataset, $\alpha_{\text{perm } 1..j,\ 1..i}$. This allows inferring the null distribution of the statistics simultaneously for all the anatomical structures at once. A two-tailed p-value was obtained by counting the number of replicates above or below the observed value in our analysis divided by the total number of replicates (i.e., 1,000) and multiplying this value by 2.

Table 2.8 Summary of the permutation test performed in the spatial embryo morphology analysis. The number of genes in the *Null distribution* column correspond to the ones analyzable with 4-fold sites.

Analysis	Hypothesis tested	Null distribution
Anatomical terms	For a given embryo anatomical term, are the genes expressed in one anatomical term undergoing differential selection compared to the genes not expressed in such embryo anatomical term?	Set of the genes expressed in the anatomical terms (4,945 genes)
Anatomical terms during embryo development	For all embryo development stages, are the genes expressed in one given anatomical term undergoing differential selection compared to the genes not expressed in that anatomical term?	Set of the genes expressed in the anatomical terms in all stages (4,943 genes).
Germ layers	For a given germ layer, are the genes expressed in one germ layer undergoing differential selection compared to the genes not expressed in such germ layer?	Set of genes expressed in the three germ layers (1,677 genes).
Germ layers during embryo development	For all embryo development stages, are the genes expressed in a germ layer undergoing differential selection compared to the other genes not expressed in a such germ layer	Set of genes expressed in the three germ layers in all stages (2,510 genes).

Chapter 3

RESULTS

3 Results

3.1. Population genomics at the DNA variation level

The first natural step in every population genomics analysis is to describe the DNA variation at the molecular level. This is accomplished by estimating the population statistics that capture the evolutionary properties of the genome sequences.

At the DNA variation level, the genome-wide variation of *D. melanogaster* was analyzed. Specifically, the proportion of substitutions that are adaptive, α, was estimated using five different approaches derived from the McDonald and Kreitman test (MKT, see Methodology, section 2.2.1). The comparison of the methodologies was performed using both empirical and simulated data to assess their statistical properties, including the number of genes necessary to conduct an MKT, the estimation of other selective regimes in addition to adaptive selection and the effect of slightly deleterious substitutions in the estimations of α.

3.1.1. Genome-wide distribution of synonymous and non-synonymous polymorphic sites and fixed differences in *D. melanogaster*

We start with a global analysis of the polymorphism and divergence levels of a total of 13,753 protein-coding genes of a North American population of *D. melanogaster* (Huang et al., 2014; Lack et al., 2015, 2016). The distribution of the total number of synonymous (at 4-fold degenerated sites, *S*) and non-synonymous (at 0-fold degenerated sites, *N*) polymor-

phic sites (P) and divergent sites (D) for the set of genes is presented in the Figure 3.1.

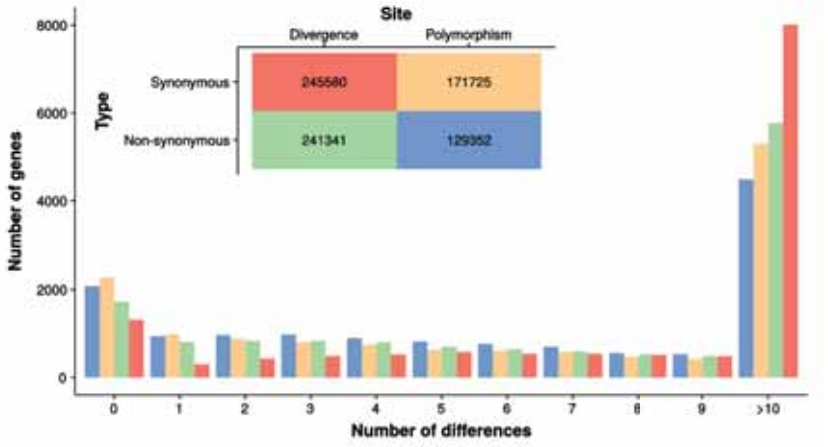


Figure 3.1 Distribution of synonymous and non-synonymous polymorphic sites and fixed differences in a North American *D. melanogaster* population. *D. simulans* is used as the outgroup species. The total megabases (Mb) analyzed are 12.10 for 0-fold and 2.91 for 4-fold degenerated sites.

12,609 protein-coding genes (91.68%) of the *D. melanogaster* population exhibit some form of coding nucleotide variation either in polymorphism or in divergence, relative to the outgroup *D. simulans*. With a total of 245,580 synonymous fixations (D_S) in 2.91 megabases (Mb) of synonymous coding region analyzed (m_S), the genomic average synonymous divergence (d_S) is 8.43%. In the case of the average non-synonymous divergence, 241,341 non-synonymous fixations (D_N) were found in 12.10 Mb of non-synonymous coding sites (m_N), hence d_N is 1.99%. For polymorphism, 171,725 synonymous polymorphisms (P_S) and 129,352 non-synonymous polymorphisms (P_N) were detected, yielding an average synonymous SNP density (p_s) of 5.9% and a non-synonymous SNP density (p_n) of 1.07%. Under neutrality, the ratios of polymorphism and divergence are expected to be equal (neutrality index, NI = $\frac{P_n / P_S}{D_N / D_S}$, Rand and Kann, 1996). However, the obtained NI is 0.77, indicating that there is an excess of non-synonymous substitutions (D_N is higher than expected under neutrality), suggesting that positive selection is a major force shaping the *Drosophila* genome (Pearson's χ^2 test with Yates' continuity correction = 3,254.3, p-value = 2.2×10^{-16}).

The same analysis was performed using *D. yakuba* as an outgroup species. In this case, the d_S increased from 8.43% to 17.87%, expected value since the divergence time is around twice in the *D. melanogaster–D. yakuba* branch (7.4 million years ago (Mya), Tamura, Subramanian, and Kumar, 2004) than in the *D. melanogaster–D. simulans* branch (up to 4.3 Mya, Cutter, 2008). In the case of non-synonymous fixations, the percentage of d_N increased from 1.99% to 3.18%. The NI, in contrast, indicates an excess of amino acid polymorphism (D_N is lower than expected), suggesting that negative selection acts to remove deleterious mutations (NI = 1.06, Pearson's χ^2 test with Yates' continuity correction = 171.95, p-value = 2.2×10^{-16}). It could also indicate that balancing selection is a common force to maintain the polymorphism in one or both species.

For the next analyses, *D. simulans* will be used as outgroup species unless the contrary is stated. The suitability of choosing one species or another as an outgroup is thoroughly discussed in the Discussion, section 4.1.1.

3.1.2. Estimation of the fraction of adaptive substitutions (α) with MKT-based approaches

The fraction of substitutions fixed by positive selection, α, was estimated using five different McDonald and Kreitman derived methodologies in the *Drosophila* protein-coding genes described above.

The first implemented approach is the standard McDonald and Kreitman test (standard MKT, McDonald and Kreitman, 1991), described in Methods, section 2.2.1. Briefly, the standard MKT assumes that positive selection can be detected as the excess of divergence relative to polymorphism at putatively selected sites. This can be quantified by the index $\alpha_{standard}$:

$$\alpha_{standard} = 1 - \frac{D_S}{D_N}\frac{P_N}{P_S} \tag{3.1}$$

One of the major limitations of the standard MKT is that it assumes that deleterious mutations do not contribute to polymorphism, contrary to what is observed in many species (Charlesworth and Eyre-Walker, 2008). In the presence of weakly deleterious polymorphisms segregating in the population, α values are downwardly biased, because P_N will be inflated.

One way to overcome this problem is to remove low-frequency polymorphisms from the analysis, which are expected to be enriched in slightly deleterious variants. This approach is known as the Fay-Wycoff-Wu method (FWW method, Fay, Wyckoff, and Wu, 2001). With the FWW method, α is estimated as in Equation 3.1 but considering only polymorphic sites with a frequency above the established cutoff, typically set at 5%, for both neutral and selected classes.

$$\alpha_{FWW} = 1 - \frac{D_S}{D_N}\frac{P_{N>5\%}}{P_{S>5\%}} \tag{3.2}$$

However, the FWW method is still expected to lead downwardly biased estimates of α due to the segregation of slightly deleterious mutations (MAF>5%) and will provide reasonably accurate estimates of α only if the rate of adaptive evolution is high and the distribution of fitness effects for slightly deleterious mutations is very leptokurtic (Charlesworth and Eyre-Walker, 2008). The higher the cutoff to remove low-frequency polymorphisms, the more minimized is the bias, but it leads, on the other hand, to a small amount of remaining polymorphism data, which lowers the statistical power to detect adaptation. To illustrate this, the expected polymorphism that neutrally segregates for a sample of 100 haploid individuals under the standard neutral model with infinite sites mutation is given by the following expression (Nielsen and Slatkin, 2013) and represented in Figure 3.2:

$$E[f_j] = \frac{1/j}{\sum_{k=1}^{n-1} 1/k}, j = 1, 2, ..., n-1 \tag{3.3}$$

Results show that 44.10% of all the polymorphisms are at a frequency <5% and would be eliminated by the FWW method (Figure 3.2).

Mackay et al. (2012) introduced the extended MKT (eMKT) to correct for the effect of slightly deleterious mutations but without losing polymorphism data as the FWW. Instead of removing all low-frequency polymorphisms under a given threshold, non-synonymous segregating sites (P_N) are split into neutral and weakly deleterious variants, and only the later are removed, thus increasing the amount of data analyzed and, as a consequence, the power of detecting selection. The formula is as:

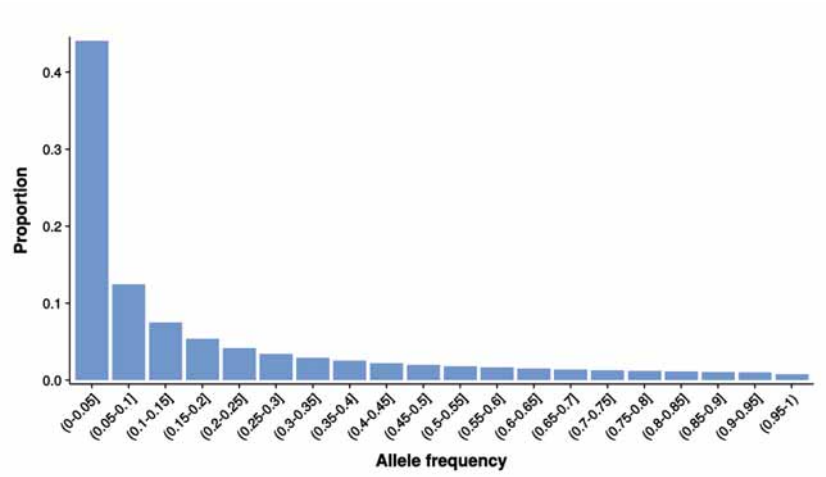


Figure 3.2 Expected site frequency spectrum (SFS) for a sample of $n = 100$ haploid individuals under the standard neutral coalescence model assuming an infinite site mutation model. Approximately half of the neutral polymorphism are at a MAF below a 10%.

$$\alpha_{extended} = 1 - \frac{D_S}{D_N}\frac{P_{N\,neutral}}{P_S} \tag{3.4}$$

One advantage of this method is that it not only removes the effect of negative selection in the estimation of α, but it provides a measure to quantify it (Mackay et al., 2012).

Finally, Messer and Petrov (2013), introduced the asymptotic MKT, a heuristic approach that takes into account the effect of slightly deleterious mutations segregating at any frequency, and not only below the 5% threshold as the previous two correction methods. First, it estimates α for each derived allele frequency (DAF) category, and then it fits an exponential function (or a linear one, when the exponential is not possible) to the values, of the form:

$$\alpha_{fit}(x) = a + b^{-cx} \tag{3.5}$$

The asymptotic α estimate is obtained by extrapolating the value of this function to 1.

$$\alpha_{asymptotic} = \alpha_{fit}(x = 1) \tag{3.6}$$

The last approach that is presented here is an extension of the asymptotic MKT, that incorporates the methodology by Mackay et al. (2012) to also calculate the fraction of sites that are under negative selection regimens. This is called the integrative MKT (iMKT). The main distinctiveness compared to the asymptotic MKT is that it calculates the negative fraction and for that, only exponential fittings are used, while asymptotic MKT incorporates both linear and exponential fittings.

Table 3.1 shows the α values estimated using the five methodologies explained above.

Standard MKT is the methodology that allows estimating α_{st} in a larger number of genes, because only genes without either synonymous or non-synonymous variable sites were not analyzed. A total of 10,505 (76.38%) *D. melanogaster* genes fulfilled this criterion. Regarding positive and negative selection, as determined with a significant and positive or a negative α_{st} at p-value < 0.05 with a Fisher's exact test, twice as many genes are detected under negative selection (1,215 genes) than under positive selection (689 genes).

When the FWW correction (at 5%) is used, a noticeable decrease in the genes that can be analyzed is observed (from 10,505 to 8,315 genes). This is because there is a significant loss of data with this method and genes cannot longer be analyzed. However, the absolute number of genes under positive selection increases an 18.14%, from 689 to 814. Finally, a noticeable drop in the number of genes under negative selection is observed in *D. melanogaster*, 85.68% less compared to the results in the standard MKT, from 1,215 to 174.

A significant increase in positively selected genes is achieved with the eMKT method (44.41% more, from 689 to 995), and also, very few genes are lost compared to the standard MKT (4.06%, 10,078 out of 10,505 can be analyzed). This method allows detecting more signals of positive selection. However, the drop in negatively selected genes is not as pronounced as with the FWW correction (from 1,215 to 670 genes), which could indicate that this method does not efficiently remove the excess of slightly deleterious polymorphism. A plausible explanation is that 4-

Table 3.1 Summary of averaged values of α, ±SD and number of analyzed genes using the five MKT approaches applied to *D. melanogaster* protein-coding genes of the Freeze 2.

Species	Set	Standard MKT		FWW method (5%)		eMKT method (5%)		Asymptotic MKT		iMKT	
		α_{st} (±SD)	N	α_{FWW} (±SD)	N	α_{ext} (±SD)	N	α_{asym} (±SD)	N	α_{iMKT} (±SD)	N
D. melanogaster (RAL)	All genes	-1.383 (±4.017)	10,505	-0.498 (±2.765)	8,315	-0.894 (±3.737)	10,078	3.212 (±11.572)	5,522	0.734 (±0.546)	237
	Positive	0.774 (±0.131)	689	0.882 (±0.111)	814	0.800 (±0.122)	995	0.861 (±0.143)	114	0.774 (±0.198)	35
	Negative	-7.147 (±8.912)	1,215	-10.094 (±11.943)	174	-8.300 (±10.833)	670	-4.441 (±3.698)	5	-	0

For standard, FWW and eMKT approaches, the *x* cutoff interval is [0,1] and a DAF of 20 categories was used. For asymptotic and iMKT approaches, *x* cutoff interval is [0,0.9] and a DAF of 10 categories was used.
Only genes with variation in P_N, P_S, D_N, D_S were analyzed, also after the FWW and eMKT frequency cutoffs were set. In eMKT, $P_{n\,neutral}$ must be higher than 0 as well.

fold degenerated sites are not completely neutral and as a consequence, less polymorphism at non-synonymous sites is removed than it should.

Asymptotic approaches, both asymptotic MKT and iMKT, are the ones that performed the worst regarding the number of genes that can be analyzed when using single-gene data. In *D. melanogaster*, some genes can be analyzed using the asymptotic approaches (5,522 genes with the asymptotic MKT and 237 with the iMKT, this latter decrease is because the exponential fit needs more data). However, a small proportion of positively selected genes can be detected with either approximation, only 114 genes can be detected as under significant positive selection with the asymptotic MKT, while only 35 genes with the iMKT.

In general, eMKT correction is the method that performed the best in terms of detecting evidence of both positive and negative selection, as it can maintain a reasonably good statistical power. For species with low-frequency variants, the FWW method is very penalizing and the loss of power is very dramatic, although it can perform well in species with a sufficient amount of data (like *D. melanogaster*). Asymptotic approaches are useless in dealing with single-gene data. In those cases, gene concatenation is a good alternative to overcome this limitation and is going to be further discussed in the following section 3.1.3.

3.1.3. Concatenating genes for estimating α

Gene concatenation is the process of merging the nucleotide variation of multiple genes into a single entity. This process has the advantage of increasing the number of polymorphic sites to construct the site frequency spectrum (SFS), and thus, gaining statistical power to implement asymptotic approaches. One important consideration is which is the optimum number of genes that should be concatenated to obtain a concatenated fragment with enough segregating sites to gain power to detect selection but not to dilute the heterogeneous behavior of different genes.

To assess which is this minimum number to obtain a representative measure of the average α of a sample of genes, a simulation was performed. 1,000 random *D. melanogaster* protein-coding genes were picked from the dataset. These genes were merged to obtain concatenated fragments of 1, 2, 5, 10, 25, 50, 75, 100, 250, 750 and 1,000 genes by resampling them 1,000 times with replacement (except for concatenated fragments of 2

and 5 genes, in which data was resampled 3,500 and 2,000 times, respectively, to ensure the representation of all the 1,000 genes), and estimated α under the iMKT method in each concatenated fragment. A DAF of 20 frequency categories was used, with an x cutoff interval of [0,0.9] (thus, removing polymorphic sites above a frequency of 0.9).

In Figure 3.3 results are shown. It is observed that only when $\leq$10 are concatenated, the average α is higher than when >10 genes are concatenated. This is because the former subset of genes is not an aleatory sample of the total 1,000 genes. Only those genes with enough polymorphic and divergent sites can be analyzed with the iMKT, which precisely are the ones with more statistical power to detect positive selection. From 25 concatenated genes onwards, the α mean stabilizes, and iMKT can estimate α in all concatenated fragments when $\geq$250 genes are concatenated. However, estimations with concatenated fragments of 75 and 100 genes already allows obtaining an accurate mean of α.

The number of genes analyzable in each concatenated fragment of this sample, the α values and $\pm$SD are shown in Table B.13.

Effect of recombination on α estimates

Recombination has an important effect on the rate of adaptive evolution (Castellano et al., 2015) as it has been explained in the Introduction, section 1.1.3. The efficiency of natural selection is expected to be maximized in high recombining regions. Because of that, an analysis categorizing the genes in concatenated fragments depending on their recombination context was performed. Recombination rates for *D. melanogaster* were retrieved from Comeron, Ratnappan, and Bailin (2012). Genes were divided into five equally sized groups depending on their recombination rates. Then, genes in each group were resampled 100 times with replacement and α was estimated in each concatenated bin using the five MKT methods. Figure 3.4 shows the results of the α values estimated with each methodology in each recombination rate concatenated fragment. Table B.14 contains the α and $\pm$SD estimated in each bin.

The amount of concatenated genes in these cases is enough for the iMKT to estimate α in most concatenated fragments (more than 2,500 genes). Results show that iMKT is the method that achieves the highest α values when a sufficient amount of data is available. An exponential increase

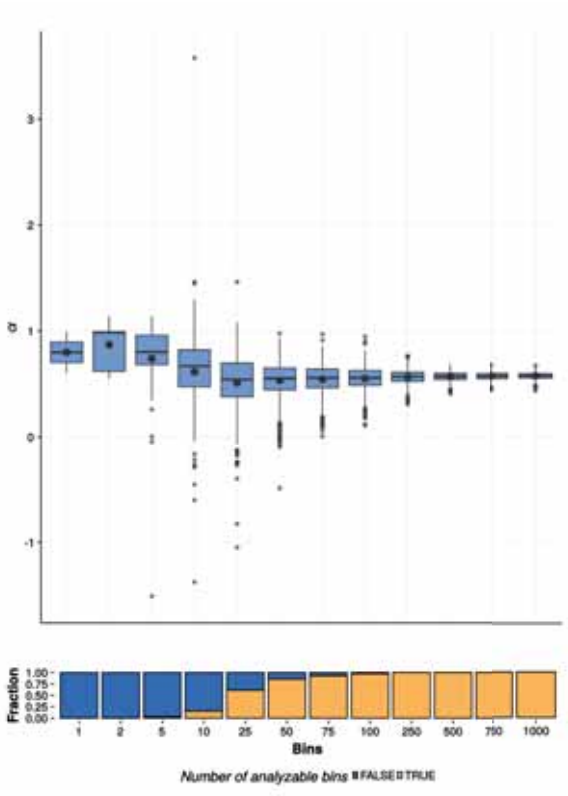


Figure 3.3 α estimated in concatenated gene fragments with the iMKT method. Each concatenated fragment was obtained by resampling 1,000 times the 1,000 random *D. melanogaster* genes with replacement, in concatenated fragments (bins) of 1, 2, 5, 10, 25, 50, 75, 100, 250, 750 and 1,000 genes (except for 2- and 5-size concatenated fragments, where data was resampled 3,500 and 2,000 times, respectively), and α was estimated in each concatenated fragment. A DAF of 20 categories was used, with an *x* cutoff interval of [0,0.9]. The bottom part of the plot represents the proportion of bins that can be analyzed (in orange) or not (in blue).

of the adaptation levels with the recombination is observed, up to a recombination level of around 2cM/Mb (as observed in Castellano et al., 2015). This plateau has been interpreted as the maximum asymptotic levels of adaptation that occur when there is no effective Hill-Robertson interference against selective sites (Castellano et al., 2015). It is worth mentioning that, under these conditions, there are no substantial differences between the FWW and the iMKT methods, especially in higher recombination rates, while the eMKT method fails by exhaustively correcting for the slightly deleterious mutations, returning downwardly biased α values. The method that performs the worst was the standard

MKT because it is the only method that does not correct for the effect of slightly deleterious mutations.

3.1.4. Testing methodologies with simulated data

Until now, the best approach was considered to be the one estimating the highest α. However, it can be the case that the α estimates are biased. For assessing a potential over or underestimation of α, the different MKT methodologies were tested on simulated data, using the forward simulation framework SLiM 2 (Haller and Messer, 2017), applying the SLiM configuration script provided in the asymptoticMK's GitHub repository (`https://github.com/MesserLab/asymptoticMK`; last accession: December 2017). The main advantage of using simulated data is that the "real" α is known. Briefly, a population of 1,000 diploid individuals evolving in 13 different scenarios was simulated, with 50 replicates for each scenario. See Methods (section 2.2.1) for a detailed description of each scenario.

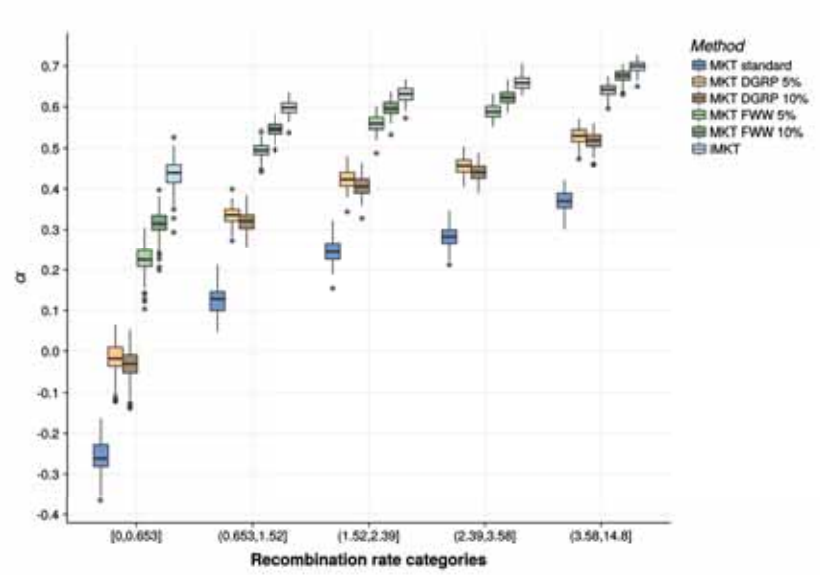


Figure 3.4 α estimated in concatenated gene fragments categorized by their recombination rate (cM/Mb) using different MKT methods. Genes were divided into five equally sized groups according to their recombination levels. Genes in each category were resampled with replacement 100 times and α was estimated in each of this 100 bins. A DAF of 20 bins was used with a x cutoff of [0,0.9] in all methods.

α was estimated in each conducted simulation using the five MKT approaches. The results from the analysis of the SLiM 2 simulations are shown in Table 3.2 and in Figure 3.5. Additionally, Table B.15 shows the mean estimation error of the different MKT approaches compared to the true α. Note that an x cutoff of [0.1,0.9] was used instead of the one of [0,0.9] that was being used in previous analyses in order to reproduce Haller and Messer's work (2017). Because of that, eMKT and FWW methods use a more stringent cutoff, removing the polymorphism below a frequency of the 15%.

Table 3.2 Results from the MKT methods for simulation runs with SLiM 2.

Simulation	Standard MKT	FWW 15%	eMKT 15%	iMKT	True α
Baseline	0.179 (±0.038)	0.207 (±0.038)	0.202 (±0.038)	0.313 (±0.05)	0.326 (±0.016)
L=10^6	0.168 (±0.104)	0.193 (±0.099)	0.189 (±0.1)	0.289 (±0.088)	0.306 (±0.07)
L=10^8	0.175 (±0.011)	0.202 (±0.011)	0.197 (±0.011)	0.297 (±0.016)	0.325 (±0.005)
T=2×10^4	0.187 (±0.103)	0.22 (±0.105)	0.214 (±0.105)	0.376 (±0.155)	0.374 (±0.076)
T=2×10^6	0.175 (±0.01)	0.203 (±0.009)	0.198 (±0.009)	0.3 (±0.014)	0.322 (±0.006)
μ=10^{-8}	0.175 (±0.011)	0.202 (±0.011)	0.196 (±0.011)	0.283 (±0.016)	0.304 (±0.005)
μ=10^{-10}	0.154 (±0.104)	0.182 (±0.103)	0.177 (±0.103)	0.299 (±0.13)	0.321 (±0.071)
s_d=0.002	0.082 (±0.04)	0.107 (±0.038)	0.103 (±0.039)	0.211 (±0.045)	0.241 (±0.014)
s_d=0.200	0.31 (±0.038)	0.335 (±0.038)	0.33 (±0.038)	0.41 (±0.047)	0.437 (±0.023)
r_b=0.0001	-0.117 (±0.054)	-0.079 (±0.05)	-0.086 (±0.051)	0.065 (±0.053)	0.085 (±0.013)
r_b=0.0010	0.382 (±0.028)	0.401 (±0.027)	0.397 (±0.027)	0.468 (±0.037)	0.491 (±0.016)
s_b=0.02	-0.112 (±0.055)	-0.072 (±0.052)	-0.08 (±0.053)	0.077 (±0.056)	0.094 (±0.015)
s_b=0.20	0.358 (±0.029)	0.38 (±0.029)	0.376 (±0.029)	0.463 (±0.045)	0.471 (±0.02)

The first row shows the average results with their ±SD of 50 replicate runs of the baseline SLiM 2 model provided in Haller and Messer, 2017's GitHub. These runs used parameter values of mutation rate μ=10^{-9} per base position per generation, chromosome length L=10^7 base positions, beneficial mutation rate r_b=0.0005, beneficial mutation selection coefficient s_b=0.1, deleterious mutation selection coefficient s_d=-0.02, and time after burn-in T=2×10^5 generations. Each subsequent row shows the results from 50 replicate runs using the non-baseline parameter value shown, while keeping the rest of values as in the baseline. True α specifies the true value of α averaged across the 50 replicates in each row; the rest represent the α values estimated from the different MKT methods. The x cutoff used in all methods is [0.1,0.9], as in the original paper (Haller and Messer, 2017). In FWW and eMKT corrections, a cutoff of the 15% was used. In all analyses a DAF of 20 bins was used.

In 13 out of 13 simulations, the mean α was lower than the real α for the standard MKT, FWW and eMKT corrections. For iMKT, only in the scenario with the smallest number of generations simulated (T=10×10^4) the mean α was higher than the real α, as observed in Haller and Messer (2017). In general, it can be observed that those simulation scenarios which created less polymorphism (i.e., a shorter simulated genomic region, lower mutation rate or a lower generation time), are the ones in which the mean estimation error of the iMKT is higher. On the contrary, simulations that produced more polymorphism provided more accurate iMKT α estimates (mean errors below 0.03, Table B.15) and they are also the ones in which the 100% of the simulations could be estimated

with the exponential fit. This was already pointed out in Haller and Messer's work (2017). A clear pattern cannot be found for the other MKT methodologies. However, it seems that the scenarios that produce a higher number of beneficial mutations (a high beneficial mutation rate, r_b or high beneficial selection coefficient, s_b) are the ones with the lowest estimation errors, both for the FWW and eMKT corrections (mean errors are < 0.1). Overall, iMKT is the method that performs the best when the level of polymorphism is high, as it is the case for these simulations. The eMKT and FWW approaches cannot correct for the presence of deleterious mutations as efficiently as the iMKT approach, and underestimate α in most of the cases. However, in cases with a low polymorphism level, closer to what is found in real data, iMKT only worked in approximately 50% of the cases and performed similarly to eMKT and FWW corrections (Figure 3.5). However, it should be taken into account that a DAF of 20 frequency bins was used and iMKT could work better by using a less fractionated DAF.

3.1.5. A flowchart to select an MKT approach

Taking into account the previous results on each presented MKT approaches, Figure 3.6 displays a flowchart that recommends the use of each methodology depending on one's data and needs. The first factor to consider is whether one wants to estimate α in individual genes or in concatenated fragments. If single gene data is used, asymptotic methods are not applicable and one is recommended to use either the eMKT, FWW or standard MKT methodology. The use of eMKT, FWW or standard MKT will depend on whether negative selection wants to be removed and/or wants to be quantified. If negative selection is not affecting the data one may use the standard MKT. This is the case of a leptokurtic distribution of deleterious effects (DFE) because leptokurtic distributions have a smaller proportion of polymorphisms that are slightly deleterious (Eyre-Walker and Keightley, 2007). On the contrary, both FWW or eMKT removes negative selection, but only the latter allows additionally its quantification and also it does not lose as much information as FWW.

If the input data is concatenated gene data, for a given threshold (Figure 3.3) asymptotic methods are preferred. For estimation of the negative fraction, the iMKT is the preferred method. Otherwise, one may use the asymptotic MKT. It has been already shown that these methods need

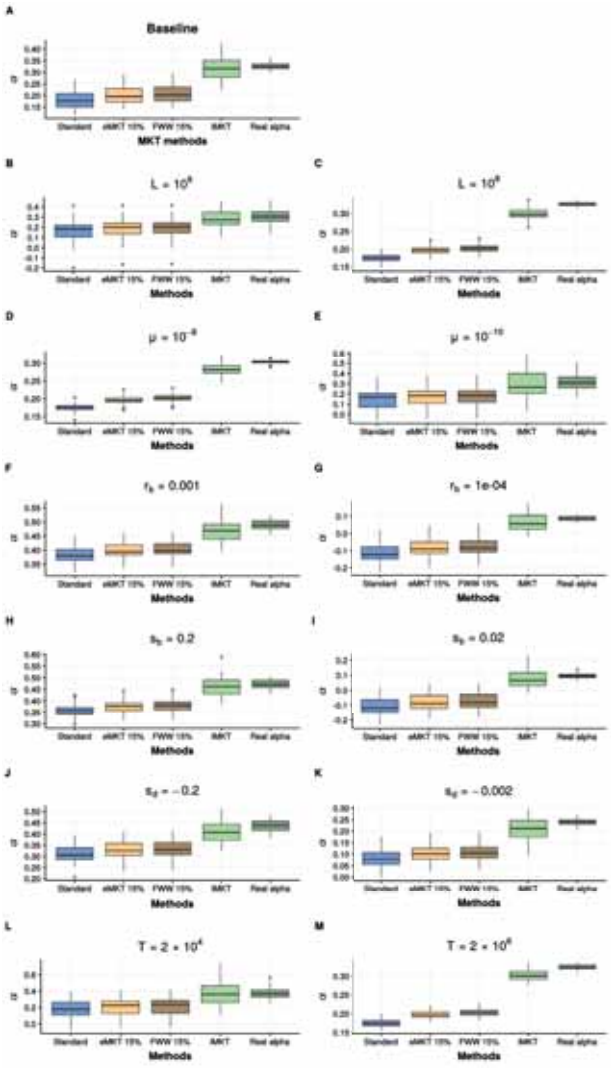


Figure 3.5 Results from the five MKT approaches for 13 simulation runs conducted with SLiM 2. A. Shows the averaged results from 50 replicate runs of the baseline SLiM model supplied on Messer & Petrov GitHub (see Methods, section 2.2.1). These runs used parameter values of mutation rate $\mu = 10^{-9}$ per base position per generation, chromosome length $L = 10^7$ base positions, beneficial mutation rate $r_b = 0.0005$, beneficial mutation selection coefficient $s_b = 0.1$, deleterious mutation selection coefficient $s_d = -0.02$, and time after burn-in $T = 2\times10^5$ generations. The subsequent graphs (**B-M**) shows the results from 50 replicate runs using the non-baseline parameter value shown in the graph title. A DAF 20 was used, with an x cutoff of [0.1,0.9].

a considerable amount of data to work. If the fitting is not possible, it is advisable to use a less fractioned DAF frequency (e.g., divide the polymorphic sites according to their frequency 10 instead of 20 equal width frequency bins). If the fitting is still not possible, then the use of the eMKT methodology is advisable for the same reasons explained above.

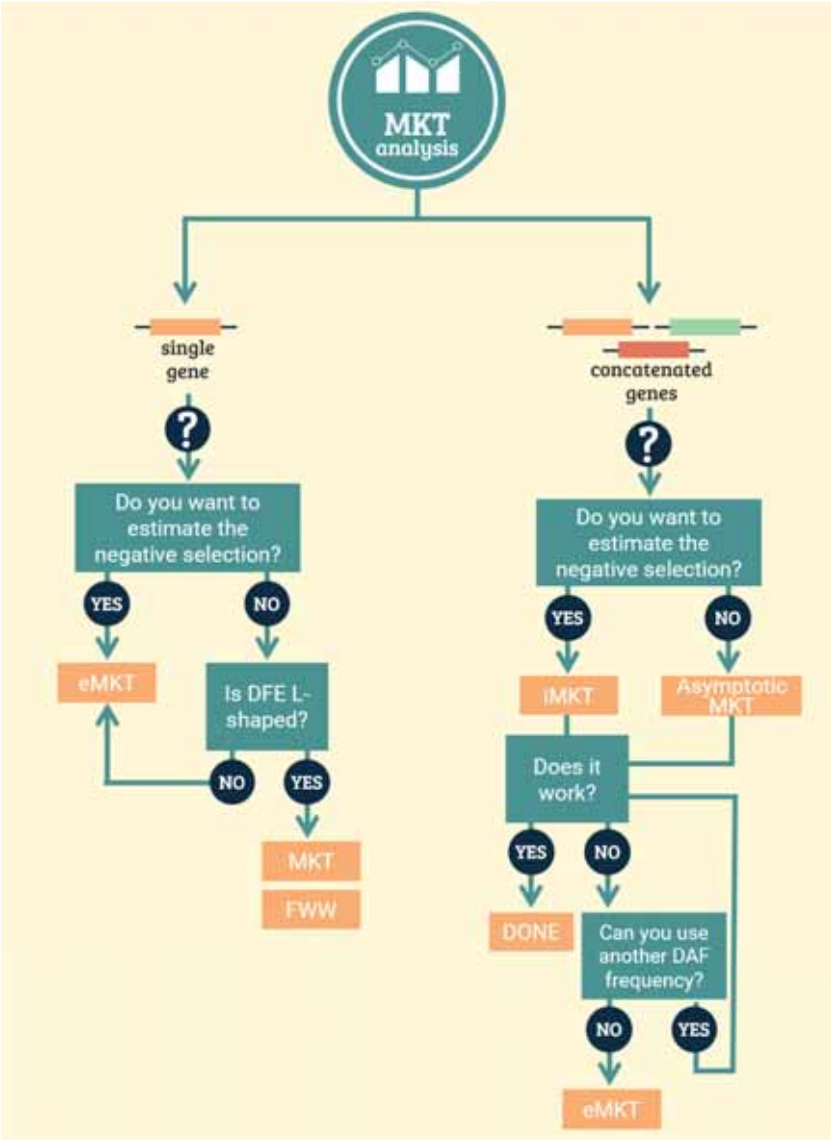


Figure 3.6 iMKT analysis flowchart.

3.1.6. Adaptation in the *D. melanogaster* genome

After the assessment of the best method to quantify both adaptation and negative selection in *D. melanogaster* protein-coding genes –the eMKT– we analyzed the general properties and functions of such genes under positive and negative selection. For that, it was obtained the list of genes that appeared to be under positive and negative selection as determined by the eMKT methodology correcting with a 5% cutoff (Table 3.1). A total of 995 genes under positive selection were detected, with an α mean of 0.8 ($\pm$0.122), and 670 under strong negative selection, with an α mean of -8.30 ($\pm$10.83). A gene ontology (GO) enrichment analysis was performed within the category *biological process* using the PANTHER-GO enrichment analysis tool (Mi et al., 2017). Table 3.3 shows the enriched GO terms for the 995 genes under positive selection with a false discovery rate (FDR) < 0.05. Among the enriched GO terms, the ones related to the immune system and sperm-related genes are remarkable. In the 670 genes under strong negative selection, no apparent enrichment of GO terms was found.

Table 3.3 Top 20 enriched GO terms in the biological process category among the *D. melanogaster* population positively selected genes.

GO biological process	Fold enrichment	FDR *p*-value
Cell-cell adhesion mediated by cadherin (GO:0044331)	7	0.021
Calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules (GO:0016339)	6	0.010
Homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156)	5.76	0.001
Male genitalia development (GO:0030539)	4.87	0.043
Homeostasis of number of cells (GO:0048872)	4.34	0.041
Regulation of hemocyte differentiation (GO:0045610)	4.34	0.040
Response to mechanical stimulus (GO:0009612)	3.67	0.038
Lipid transport (GO:0006869)	3.25	0.035
Multicellular organismal homeostasis (GO:0048871)	2.92	0.036
Heart development (GO:0007507)	2.67	0.012
Dorsal closure (GO:0007391)	2.67	0.017
Digestive tract development (GO:0048565)	2.56	0.040
Open tracheal system development (GO:0007424)	2.25	0.003
Developmental growth (GO:0048589)	2.18	0.025
Immune response (GO:0006955)	2.11	0.026
Imaginal disc development (GO:0007444)	2	3.2×10^{-5}
Compound eye development (GO:0048749)	1.89	0.010
Regulation of organelle organization (GO:0033043)	1.87	0.009
Epithelial cell differentiation (GO:0030855)	1.85	0.007
Positive regulation of transcription, DNA-templated (GO:0045893)	1.82	0.016

3.2. Population genomics at the genomic level

The population genomic statistics estimated on protein-coding genes are correlated with other genomic features to evaluate the impact of these properties on the nucleotide variation of these genes. The aim is to discover variation patterns of protein-coding genes updating previously known ones by adding new layers of molecular genomic information. Thanks to the availability of complete genomes and high-quality functional genomics datasets (see Table 1.2 for a compendium of *D. melanogaster* -omics resources), together with refined statistical methods, light can be shed to one of the most long-standing problems in molecular population genetics: understanding what genetic, genomic, expression and phylogenetic features governs the evolution of protein-coding genes.

3.2.1. Measured genomic features

First of all, a number of features of *D. melanogaster* were measured to relate them to the variation level of protein-coding genes. Those features span different characteristics of the genome: (i) gene architectonic features: gene size, as the length of the coding sequence of a gene; number of exons and transcripts, as the total number of exons and transcripts a gene has; intron length, measured as the average distance in base pairs between the exons of a given gene; (ii) expression features: expression bias, a measure of how evenly distributed the expression of a gene is over time; expression level, the average expression of a gene over all life cycle stages; codon usage bias, measured as the frequency of optimum codons, *Fop*; (iii) genomic context features: recombination rate, based on observed cross-overs in 100 kb intervals, from Comeron, Ratnappan, and Bailin (2012); intergenic distance, as the average distance in base pairs between two adjacent genes; (iv) phylogenetic features: phylogenetic age, using the phylostratigraphic maps from Drost (2014). A complete description of the measurement of these genomic features can be found in Methods, section 2.2.4. This information constitutes one of the most complete genomic feature datasets used for characterizing a species genome.

3.2.2. Correlation between selective regimes and features

The relationship between the set of estimated features and the selective regimes parameters was assessed. Four different parameters were estimated using DFE-alpha software (Eyre-Walker and Keightley, 2009). First, α_{DFE} (α from now onwards), the proportion of substitutions that are adaptive. Second, ω (d_N/d_S), used as a proxy for conservation at the sequence level. ω is the rate of non-synonymous substitutions per non-synonymous site (d_N) divided by an estimation of the mutation rate (d_S, the rate of synonymous substitutions per synonymous site) in a gene. ω can be subdivided into two statistics: ω_a and ω_{na}. Third, ω_a is estimated as $\omega \times \alpha$, or the rate of adaptive substitutions per non-synonymous site divided by the synonymous substitution rate (Gossmann, Keightley, and Eyre-Walker, 2012). Forth, ω_{na} is $\omega \times (1 - \alpha)$, the rate of non-adaptive substitutions (i.e., nearly neutral and deleterious, Galtier, 2016).

For analyzing the patterns of variation, genes in the dataset were categorized in five different categories (when possible) based on each genomic feature (see Table B.11 for the number of genes considered in each category) and resampled with replacement 100 times the genes in each category and estimated the selection statistics in each category bin in order to estimate the CI of each parameter (see Methods, *Gene resampling (bootstrapping)* section).

Genomic features negatively correlated with the selective regimes

The genomic features negatively correlated with the selective regimes are shown in Figure 3.7. All four analyzed features related to the architecture properties of genes are negatively correlated with the evolutionary rate of proteins, which are gene size, number of exons and transcripts and intron length.

Gene size follows a reverse J-shaped distribution, highly positively skewed and with a long tail (Figure 3.7A). It ranges from 9 to 55,400bp (a CDS of 9bp is likely due to errors in the *D. melanogaster* annotation file). Each gene size category is composed of more than 2,000 genes (Table B.11), so attributed errors to the annotation accuracy are expected to be negligible.

Gene size is linearly negative correlated with both ω_a (Figure 3.7B) and ω_{na} (Figure 3.7C). These observations indicate that the longer the gene, the less adaptive and non-adaptive substitutions become fix, i.e., purifying selection is more efficient on longer genes. On the contrary, shorter genes experience a reduction of the efficacy of natural selection, and thus, they tend to fix more adaptive and non-adaptive substitutions. These results are also found in a number of studies (see Table 1.1), but only related to ω and not to ω_a and ω_{na} as in here. This allows us to infer that the higher substitution rates experienced by short genes are due to the accumulation of both adaptive and negative selection.

The number of exons and transcripts, similarly to the gene size, follow a reverse J-shaped distribution and highly positively skewed (Figures 3.7D and 3.7G, respectively). Regarding the number of transcripts feature, genes could not be divided in equally-sized categories, due to the high proportion of genes having 3 or fewer transcripts annotated (almost half of them only have one transcript annotated) and therefore, genes were divided into 4 categories.

Both the number of exons and number of transcripts show clear negative correlations with both ω_a (Figures 3.7E and 3.7H, respectively) and ω_{na} (Figures 3.7F and 3.7I, respectively). Genes having 3 or more exons annotated experience a similar rate of adaptive substitutions; while the relationship with the rate of non-adaptive substitutions is linear. In the case of the number of transcripts, a similar trend is found. To our knowledge, only the number of exons feature has been correlated with ω (Guillén, Casillas, and Ruiz, 2018) with similar results.

The last architectonic feature is the intron length, which also follows a reverse J-shaped distribution, highly positively skewed and with a long tail (Figure 3.7J).

The correlation between intron length and the selective regimes is also negative correlated with both ω_a (Figure 3.7K) and ω_{na} (Figure 3.7L). That indicates that genes having longer introns, experience a more efficient purifying selection than genes with short or without introns. A number of studies found the same trend when analyzing d_N and ω. Marais et al. (2005) argue that this correlation can be explained by a higher abundance of *cis*-regulatory elements within introns in genes under strong purifying selection. Because of their important role during development (for a review see Spitz and Furlong, 2012), it is thoroughly discussed in section 4.2.3.

Two features related to the expression level also show a negative correlation with both ω_a and ω_{na}: the codon usage, measured as *Fop*, and the expression level.

The *Fop* follow a bell-shaped distribution and ranges from 0 (no optimal codons are used) and 1 (only optimal codons are used) (Figure 3.7M). Typically, it is used as a measure of the expression level.

The correlation between the *Fop* and the selective regimes is also linear, but the distribution follows a reverse J-shaped curve rather than a linear relation as the previous four features. That indicates that genes with a higher *Fop* are more constrained (Figure 3.7N) and fix less adaptive substitutions (Figure 3.7O). The negative correlation is expected because, as found in a number of studies (Table 1.1), the correlation is likely driven by purifying selection against mutations that reduce the transcriptional and/or translational efficiency and/or robustness of proteins. Genes with a higher adaptive rate would have fixed many amino acids that do not necessary are the optimal ones (Larracuente et al., 2008).

The last feature negatively correlated with the selective regimes is the expression level. Expression level follows a bimodal distribution and positively skewed (Figure 3.7P).

In agreement with the literature (Table 1.1), genes that are highly expressed are more constrained than low-expressed genes (Figures 3.7Q and 3.7R, respectively). Therefore, highly expressed genes fix less adaptive and non-adaptive substitutions. Some studies (Table 1.1) also find the same pattern with ω. The most plausible explanation could be that new non-synonymous mutations affecting the transcription or translation of proteins (e.g., leading to misfolding or misinteraction) will have a stronger deleterious effect in highly expressed genes than in lower ones. Additionally, a number of studies (Carneiro et al., 2012; Williamson et al., 2014; Hodgins et al., 2016) have also correlated the expression level with ω_a using DFE-alpha in mammals (European rabbit) and plants (*Capsella grandiflora*, lodgepole pine and interior spruce) finding consistent results.

In general, it can be inferred that purifying selection acts more efficiently on complex and highly expressed genes, i.e., those that are longer, containing more exons and longer introns, encoding a large number of isoforms and that are highly expressed, while the contrary happens with less complex and less expressed genes.

Genomic features positively correlated with the selective regimes

Only four features are positively correlated with the selective regimes. None of them are related to the architectonic properties of genes, but with the expression level, genomic context and phylogenetic features.

The only expression feature positively correlated with the selective regimes is the expression bias. The expression bias follows a bimodal distribution and it ranges from 0 to 1 (Figure 3.8A). A gene with an expression value close to 0 indicates that it is broadly expressed across multiple stages, while a value of 1, that is specifically expressed in a few stages.

The expression bias is positively correlated following a J-shaped curve with both ω_a (Figure 3.8B) and ω_{na} (Figure 3.8C). That indicates that natural selection is less efficient on genes that are specifically expressed in a few stages. On the contrary, genes that are ubiquitous are under constraint.

This pattern has already been found by a number of studies (Table B.11) and the most plausible explanation is that the extensive pleiotropy experienced by ubiquitously expressed genes will constraint them. Ubiquitously expressed genes may be are involved in more cellular and physiological processes than stage-specific genes.

Only one variable clearly enhance the efficacy of purifying and positive selection at the same time: the recombination rate. It follows a reverse J-shaped distribution and it is positively skewed (Figure 3.8D).

Recombination increases the independence of sites between genes, therefore, genes in high recombination environments can effectively fix adaptive substitutions (Figure 3.8E). The increase is not linear but rather asymptotic, in agreement with reaching an adaptation level that occurs when there is no effective Hill-Robertson interference against selective sites. On the contrary, the correlation with ω_{na} follows a negative asymptotic relationship (Figure 3.8F).

To some extent, this trend is also observed with the intergenic distance measure, that could have a similar effect as the recombination rate. This feature also follows a reverse J-shaped distribution and highly positively skewed (Figure 3.8J). More isolated genes tend to fix more adaptive substitutions (Figure 3.8K) and remove non-adaptive substitutions (Figure

3.8L). However, the correlations are not clearly linear. Contrary to what is expected (assuming that there is a linear correlation as with the recombination), genes that are close to each other or completely nested fix adaptive substitution higher than expected. On the other hand, genes that are isolated, fix a high proportion of non-adaptive substitutions.

The phylogenetic age is the last feature that we analyzed. Because its a discrete feature (can only take values from 1 to 13) rather than a continuous variable as the other ones, it was manually divided into 5 categories, trying to keep the same number of genes. Therefore, it does not follow a clear distribution 3.8G).

The correlation with the phylogenetic age and ω_a follows a J-shaped curve (Figure 3.8H). The same is found with ω_{na} (Figure 3.8I). That indicates that phylogenetically younger genes, more specifically the ones that appeared in Diptera (phylogenetic ages 12–13), accumulate more adaptive and non-adaptive substitutions than phylogenetically older ones.

All the aforementioned features are discussed in detail together with their role during development in section 4.2.

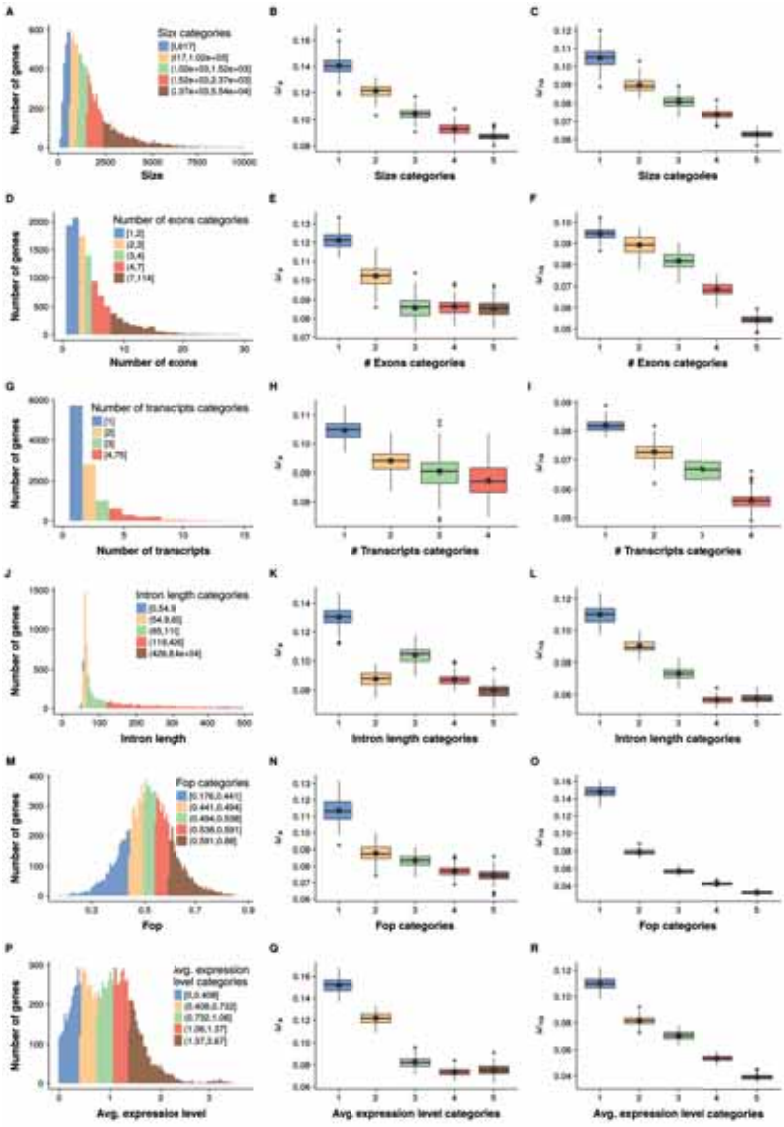


Figure 3.7 Genomic features negatively correlated with ω_a and ω_{na}. **A**. Gene size distribution and categories. **B**. Relationship between ω_a and size. **C**. Relationship between ω_{na} and size. **D**. Number of exons distribution and categories. **E**. Relationship between ω_a and number of exons. **F**. Relationship between ω_{na} and number of exons. **G**. Number of transcripts distribution and categories. **H**. Relationship between ω_a and number of transcripts. **I**. Relationship between ω_{na} and number of transcripts. **J**. Intron length distribution and categories. **K**. Relationship between ω_a and intron length. **L**. Relationship between ω_{na} and intron length. **M**. *Fop* distribution and categories. **N**. Relationship between ω_a and *Fop*. **O**. Relationship between ω_{na} and *Fop*. **P**. Expression level distribution and categories. **Q**. Relationship between ω_a and expression level. **R**. Relationship between ω_{na} and expression level. Each boxplot (100 bootstrap replicates per category) in a plot is calculated for a randomly drawn sample of the set of genes in each category with replacement. See Table B.11 for the number of genes considered in each category.

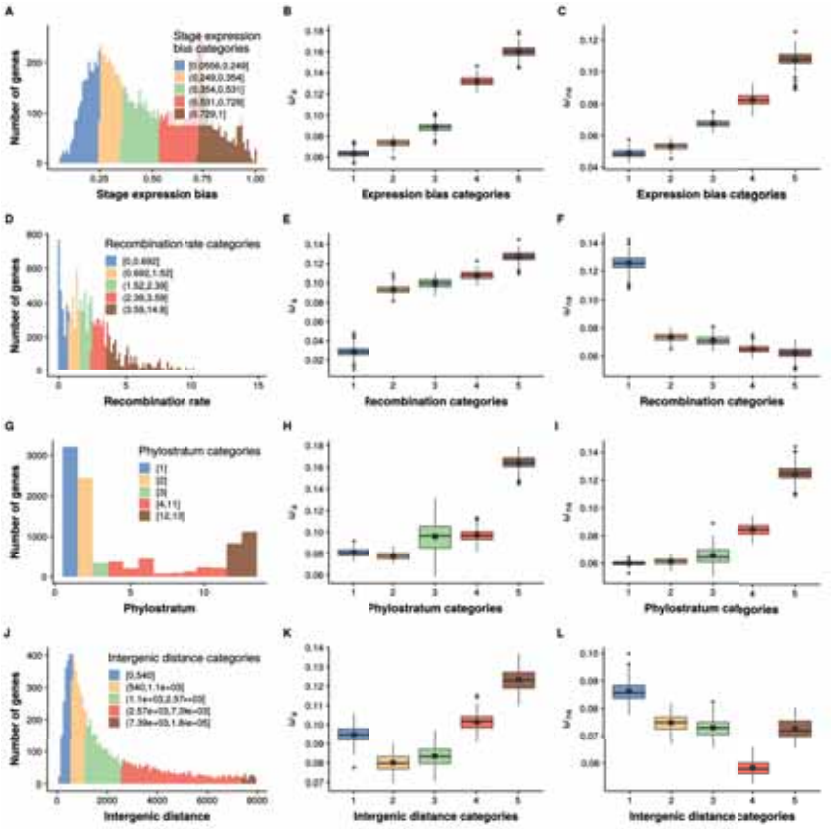


Figure 3.8 Genomic features positively correlated with ω_a and ω_{na} A. Expression bias distribution and categories. **B**. Relationship between ω_a and expression bias. **C**. Relationship between ω_{na} and expression bias. **D**. Recombination rate distribution and categories. **E**. Relationship between ω_a and recombination rate. **F**. Relationship between ω_{na} and recombination rate. **G**. Phylogenetic age distribution and categories. **H**. Relationship between ω_a and phylogenetic age. **I**. Relationship between ω_{na} and phylogenetic age.**J**. Intergenic distance distribution and categories. **K**. Relationship between ω_a and intergenic distance. **L**. Relationship between ω_{na} and intergenic distance. Each boxplot (100 bootstrap replicates per category) in a plot is calculated for a randomly drawn sample of the set of genes in each category with replacement. See Table B.11 for the number of genes considered in each category.

3.3. Population genomics at the multiomics level

The advent of NGS technologies has made available functional genomic datasets measuring features such as gene expression across different developmental stages or body parts. In this work, high quality expression data from modENCODE (Consortium et al., 2010) and BDGP (Tomancak et al., 2007) has been integrated with population genomics data to infer adaptation and selective constraint in different moments of the development and body parts.

3.3.1. Overall temporal pattern of adaptation and selective constraint over the life cycle of *D. melanogaster*

The pattern of adaptation and selective constraint is measured over the whole life cycle of *D. melanogaster*. Although different works analyze the pattern of constraint and expression divergence during development (e.g., Kalinka et al., 2010; Levin et al., 2016, see section 1.3.2), little is known about the role of natural selection in the life cycle of *D. melanogaster*. We give a global perspective on how natural selection acts during the development of this species, trying to assess whether different regimes of selection act on different developmental stages.

Four selective regimes (ω_a, ω_{na}, ω and α) have been calculated in the set of genes expressed in each life cycle stage using DFE-alpha. Additionally, a fifth statistic was also calculated, the proportion of effectively neutral mutations, P_0, based on polymorphism data alone, using the program `prop_muts_in_s_ranges.c` that comes with DFE-alpha (see section 2.2.1). Finally, the transcriptome divergence index (TDI), a measure of the average transcriptome selection pressure, was also calculated (Quint et al., 2012). The TDI is computed as the ω of each gene weighted by its relative expression in each stage (see Methods, section 2.2.4).

Figure 3.9 shows the temporal pattern found when using the DFE-alpha method. Short introns were used as a proxy for the neutral mutation rate, considering all genes with non-zero expression (and excluding the 6,655 genes that were constitutively expressed throughout all stages, see section 2.2.2 and Table B.1). A total of 2,869 genes fitted to this criterion. Both adaptive (ω_a) and non-adaptive (ω_{na}) substitution rate are the highest in the set of genes expressed at the very first embryonic

stage. Both substitution rates gradually decrease until the 10-hour embryo stage. The next developmental stages (mid and late embryonic development) show, on the contrary, the lowest substitution rates (either adaptive or not). At the third larval stage the rate of adaptive substitutions (ω_α), and to a lesser extent, the rate of non-adaptive substitutions (ω_{na}) increase and remain high through all the pupal stages. Finally, in the male adult stage, ω_α and ω values are very similar to those of the pupa while female adults exhibit lower values.

To analyze whether these differences between stages were statistically significant, stages were merged into eight developmental periods: embryo 0–2 h, embryo 2–6 h, embryo 6–24 h, larva 1–3, larva 4–6, pupa, females and males. By means of a permutation test, the probability that the genes expressed in a period undergo differential selection compared to the genes not expressed in that period was calculated, using as a null model the 2,869 genes expressed during the whole development (see Methods section 2.3.1). This analysis shows that mid and late embryonic development, the beginning of the larva and genes expressed in female adults show significantly low rates of non-synonymous substitutions (Figure A.2). The relatively high rates of substitutions in early development and in the larva, pupa and males are not significant in this analysis. *P*-values are shown in Table B.16.

However, if the same permutation test is done using the whole gene dataset (expressed in all stages or not) as the null model, the test shows that early development, late larva, pupa and male adult exhibit significantly high ω_a, ω and α values (Figure A.3). *P*-values are shown in B.17. The results of these permutation tests indicate that, in general, a vast majority of *D. melanogaster* protein-coding genes are under negative selection while a high proportion of the genes expressed in the development, under this criterion of expression, are biased toward less constrained genes.

We performed the following validity checks to confirm these results under different conditions and criteria.

TESTING FOR DIFFERENT MKT METHODOLOGIES. Three different methods to estimate α were compared since the use of different methodologies can yield different α estimates. Similar results were found when using the standard MKT and eMKT (Figure A.4) as alternatives compared to DFE-alpha. The ω_α and α statistics were, however, slightly lower for both standard MKT and eMKT than for DFE-alpha. This is especially

evident for the standard MKT, which does not correct for the effect of slightly deleterious polymorphisms.

TESTING THE PUTATIVE NEUTRAL CLASS. The estimation of the different statistics relies on a selectively neutral class of sites in the genome. Two classes of sites were used: positions 8 to 30 at short introns (as in Halligan and Keightley, 2006) and 4-fold degenerated sites (see Methods, section 2.2.1). Very similar results were found by both approaches (see Figure 3.9 for the result using short introns and Figure A.5 when using 4-fold degenerated sites). The values of ω, ω_a and ω_{na} are, overall, larger when the statistics are estimated based on 4-fold degenerated sites as neutral reference.

TESTING THE STATISTICAL POWER. The unequal number of genes expressed in each developmental stage can have an effect on the values of the estimated metrics. The pattern of Figure 3.9 was performed resampling the number of genes expressed in each stage. This analysis was repeated but resampling the same number of genes in each stage (350 genes per stage with replacement 100 times) and calculating the mean values for the selection metrics. Figure A.6 shows that very similar values are found for each selection statistic over time.

TESTING DIFFERENT EXPRESSION CRITERIA. The criterion used to consider a gene as expressed at a stage can have a major effect on the temporal patterns of the different statistics. In Figure 3.9 a gene is considered expressed at a stage if at least one transcript read (RPKM) is reported in the RNA-seq experiment in such stage –this is a *low stringent criteria*, see section 2.2.2. Very similar results were found when considering only genes that have a maximal expression level (over all stages) that is at least twice (or four times) its minimal expression level (Figure A.7). Similar results are also found if in each stage the genes that have two or more transcript reads (Figure A.8, see Methods, section 2.2.2, *Medium stringent criterion*) or ten or more Figure A.9, see Methods, section 2.2.2, *High stringent criterion*) were considered. In this case, the stages with maximum and minimum ω_a, ω_{na} and ω are the same that in the previous analyses, but the overall temporal profile is smoother. The comparison of these results with the original analyses depicted above indicates that the conservation of mid and late embryonic development (low ω and ω_{na}) is stronger in genes with high expression levels than those with low expression levels (Figures A.9C and A.9D), in agreement with previ-

ous reports of slower rates of evolutionary change in strongly expressed genes (Pal et al., 2001).

ACCOUNTING FOR THE IMMUNE SYSTEM AND TESTIS-RELATED GENES. Immune system and testis-related genes have been reported to be under higher rates of adaptation than other genes in a number of publications (Pröschel, Zhang, and Parsch, 2006; Haerty et al., 2007; Obbard et al., 2009). It was investigated whether the results can be explained by the testis and immune genes alone. Our results are roughly the same when excluding genes related with the immune system and sperm-related genes (Figure A.10), so the temporal pattern of conservation and adaptation and the high rates of substitution observed are not due to these genes.

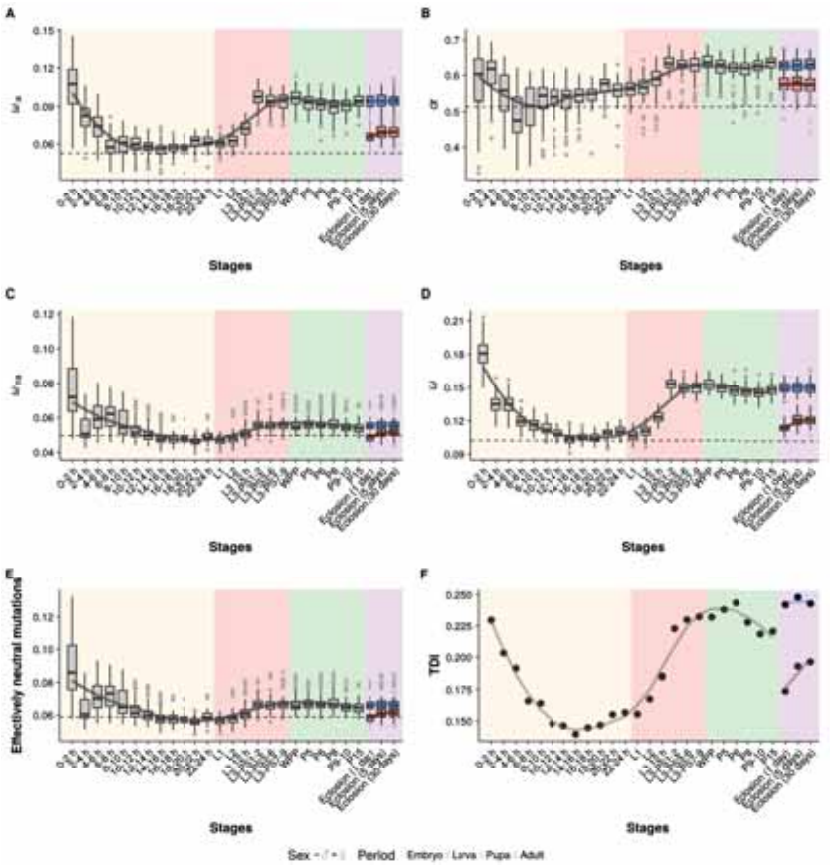


Figure 3.9 Temporal pattern of the four selective regimes indexes estimated with DFE-alpha (ω_a, α, ω_{na} and ω), P_0 and TDI. A. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **B**. α, the proportion of base substitutions fixed by natural selection. **C**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **D**. ω, the rate of non-synonymous substitutions relative to the mutation rate. **E**. Proportion of effectively neutral mutations. **F**. Transcriptome divergence index (TDI). Each boxplot (A-E, 100 bootstrap replicates per stage) in a plot is calculated for a randomly drawn sample of the set of genes expressed in a stage with replacement. The solid line going through the boxplot is inferred by LOESS. For the male and female stages the line is simply a linear regression. The dashed line shows the mean value of each statistic for the genes that are expressed in all stages (again with 100 bootstrap replicates). The TDI is the ω of each gene weighted by its relative expression in each stage (see Methods, section 2.2.4). The embryonic stages are named by the hour's intervals (from 0h to 24h), the larval stages are the first instar (L1), second instar (L2) and third instar (L3). The L3 stages are subdivided into the first 12 hours (L3-12h) and several puff stages (L3-PS1 to L3-PS7). WPP is the white pre-pupae stage. The pupal stages with RNA-seq are phanerocephalic pupa, 15h (P5), 25.6 hours pupa (P6), yellow pharate, 50.4 hours (P8), amber eye-pharate, 74.6 hours (P9-10), green meconium pharate, 96 hours (P15). Adult stages are 1, 5 and 30 days after eclosion (1 day, 5 days and 30 days). Number of genes analyzed are in Table B.1.

3.3.2. Gene expression profile clustering

There are at least three different scenarios that could explain the observed temporal pattern of change in the estimated selective regimes. In the case of ω_a for example, it could be that a subset of genes with high ω_a is expressed just with the observed temporal pattern. Alternatively, it could be that each of the time periods with high ω_a would express a distinct group of genes that have high levels of adaptive substitutions. It could also be that no simple correspondence exists between the high ω_a in a time period and the expression of a specific subset of genes in it.

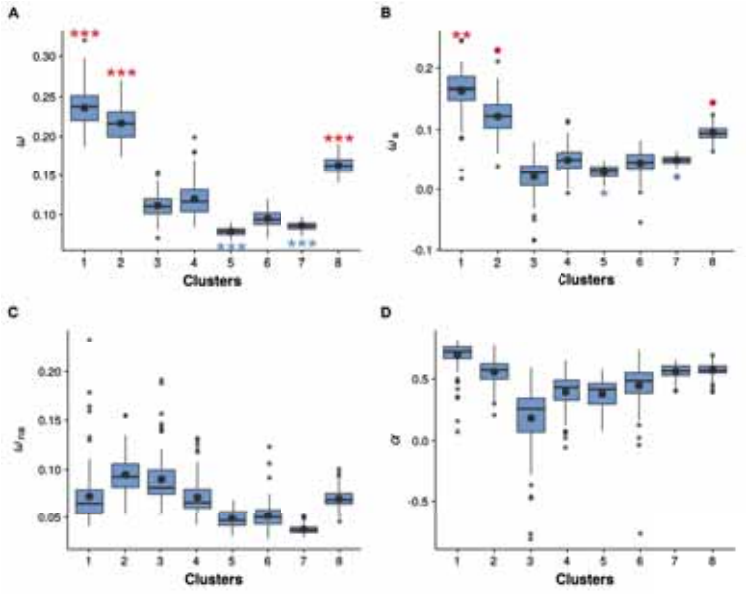


Figure 3.10 Selective regimes (ω, ω_a, ω_{na} and α) estimated using DFE-alpha over development clusters. A. ω for each cluster. **B**. ω_a for each cluster. **C**. ω_{na} for each cluster. **D**. α for each cluster. Each point in the plots is calculated for a randomly drawn sample of the set of genes in each cluster with replacement (100 bootstrap replicates per cluster). Number of genes analyzed in Table 2.3. Permutation *p*-values are shown in Table B.18.

To explore these three possibilities, all the analyzed genes were categorized into classes based on their temporal profiles of expression. To do that, an unsupervised soft clustering algorithm was used (Futschik and Carlisle, 2005) as explained in the Methods (section 2.2.2). Genes within each temporal expression class show relatively similar changes in gene expression levels over time. Eight such classes were considered for embryonic development (Figure 2.7, Table 2.3) and nine classes for the

whole life cycle (Figure A.1, Table B.9). For the embryonic development, clusters 1 and 2 are the ones showing the highest significant ω_a and ω compared to the other clusters (cluster 1: ω: p-value < 0.001; ω_a: p-value = 0.008; cluster 2: ω: p-value < 0.001; ω_a: p-value = 0.059, Figure 3.10B and 3.10E). These clusters correspond to the genes that are expressed at high levels in the earliest development and that rapidly decrease their expression to very low levels. ω_a, ω and α values in clusters 1 and 2 are larger than those in the first three developmental stages and, thus, it is likely that the genes in these clusters are responsible for the high ω_a, ω, ω_{na} and α values in the earliest development. The decline in the values of these selection statistics over early development would then just be a simple reflection of the decrease in expression of the genes in those clusters over time. Cluster 8 also shows larger ω than the ones found in the other clusters (permutation test, ω: p-value < 0.001). Cluster 8 is composed of genes whose expression increases only in the last stages of embryonic development. This high ω_a cannot be detected when directly analyzing the genes in each stage because the other genes expressed in these late stages have lower ω values, as it can be seen for cluster 5, which expresses genes from the 10 hour onwards and are constrained. Thus, the temporal pattern of change in the selection statistics seems to come from the temporal dynamics of expression of three different sets of genes (those of cluster 1, 2 and 8). Table B.18 contains the p-values. Similar results were found when the clustering was done over the whole life cycle (Figure A.11, Table B.19).

3.3.3. Genomic features correlation

The previous correlations between the population statistics (α, ω_a, ω_{na} and ω) and the genomic features (section 3.2) are extended to the developmental level by assessing how these genomic features change during time when considering the genes expressed in different developmental stages.

Genomic features exhibit a temporal pattern that either mirror that of ω_a or exhibit the opposite pattern that of ω_a (Figure 3.11). To analyze these relationships, the correlation between the average ω_a of each stage and the average of each genomic feature by stage was calculated (Table 3.4). Gene size, number of exons, *Fop* and number of transcripts per gene follow a temporal pattern that is the opposite of that of ω_a. Intron length also shows a temporal pattern opposite to that of ω_a except that no clear differences between stages are found after embryonic development. The intergenic distance shows a temporal pattern similar to that of ω_a, except that this distance is low in the earliest stages in which ω_a is high. The average expression bias shows a temporal pattern similar to that of ω_a except for an overall increase over time. The average expression level simply decreases over developmental time and the life cycle. The same correlations are found when 4-fold degenerated sites were used as a proxy for the mutation rate (Table B.20 and Figure A.12).

Table 3.4 Spearman's correlations between ω_a and genomic features.

Genomic feature	Relation with ω_a	Correlation (r^2) for females (p)	Correlation (r^2) for males (p)
Intron length	Negative	0.802 (1.12×10^{-6})	0.808 (1.08×10^{-6})
Gene size	Negative	0.731 (1.56×10^{-6})	0.764 (1.39×10^{-6})
Number of exons	Negative	0.862 (6.53×10^{-7})	0.886 (4.82×10^{-7})
Number of transcripts	Negative	0.874 (5.70×10^{-7})	0.870 (5.94×10^{-7})
Fop	Negative	0.759 (1.42×10^{-6})	0.688 (1.71×10^{-6})
Expression bias	Positive	0.508 (4.89×10^{-5})	0.552 (1.58×10^{-5})
Recombination	Positive	0.330 (2.07×10^{-3})	0.334 (1.91×10^{-3})
Intergenic distance	N.S.	0.043 (0.299)	0.082 (0.148)
Expression level	Negative	0.303 (0.003)	0.412 (4.19×10^{-4})
Phylogenetic age	Positive	0.700 (1.67×10^{-6})	0.649 (2.09×10^{-6})

See Figure A.13 for the correlations. Spearman's correlations performed between each stage's average ω_a and the average of each genomic feature in each stage. Females and males are separated because their gene expression is measured separately in the last three stages in the modENCODE.

A similar pattern is found when the genomic features were analyzed in the gene expression clusters for the embryo development and life

cycle (see Figure A.14 and *p*-values in Table B.21 and Figure A.15 and *p*-values in Table B.22, respectively). That indicates that those genes that are expressed in clusters showing selective constraint exhibit particular features. For example, cluster 5, a set of genes expressed specifically in mid-development that appear to be one of the most constrained set of genes (Figure 3.10A), expresses on average, genes that are longer, with more exons and transcripts (Figure A.14). On the contrary, cluster 8, a set of genes specifically expressed at the end of the development, is relaxed (Figure 3.10A), and the genes expressed are shorter, with short introns and fewer exons and transcripts (Figure A.14).

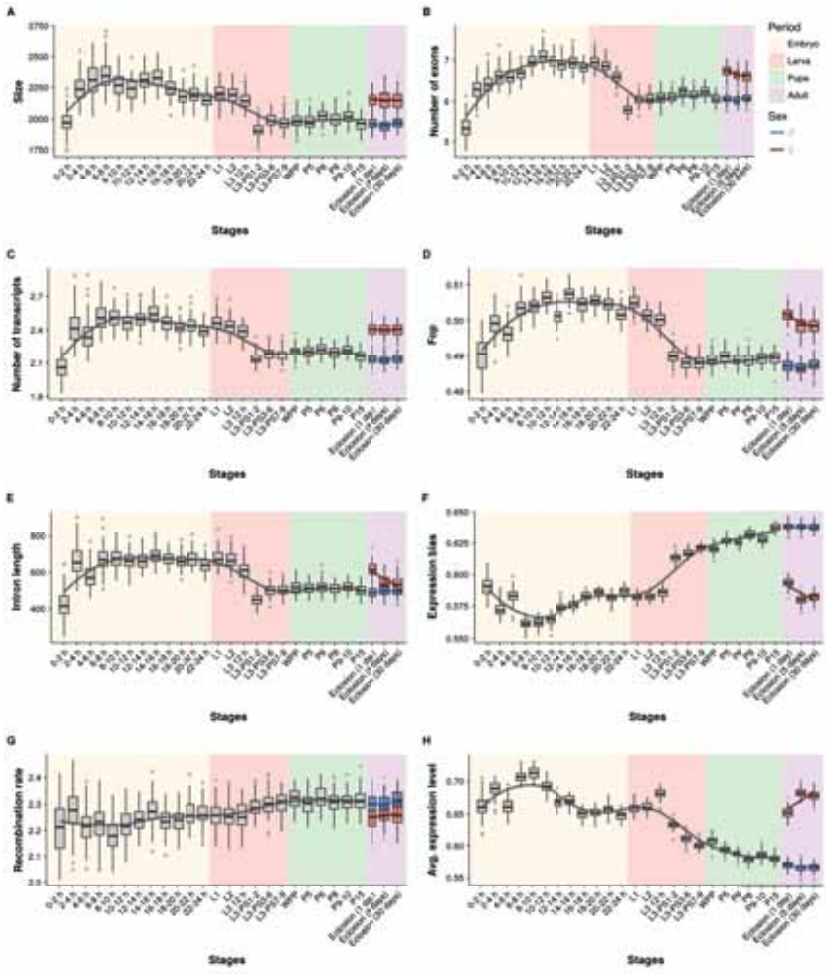


Figure 3.11 Temporal pattern of six genomic features over developmental stages. Lines and stages as in Figure 3.9. **A**. Gene size is the coding sequences length of a gene in base pairs. **B**. Exons are the number of exons for the genes expressed in a stage. **C**. Transcripts are the number of different transcripts in each gene expressed in a stage. **D**. *Fop* is a measure of codon usage bias: the ratio of optimal codons to synonymous codons. **E**. Intron length is the average distance, in bases, between the exons of a gene. **F**. The expression bias is a measure of how much the expression of a gene is restricted to one or few stages estimated as Equation 2.7 (see Methods, section 2.2.4). **G**. Recombination rate is estimated in 100 kb non-overlapping windows. **H**. Expression level is the average expression (as the logarithm of the RPKM counts) of a gene in over all stages. Mean sampling distribution was obtained by resampling 100 times with replacement the genes from each stage. See Table B.1 for the genes considered in each stage. The same patterns are found when using 4-fold data, see Figure A.12.

3.3.4. Analysis of maternal, maternal-zygotic and zygotic genes

To further explore the high ω_a and ω_{na} values in the earliest stages, maternal, maternal-zygotic and zygotic genes were analyzed separately. For that purpose, a microarray study was used (Thomsen et al., 2010). This study categorized developmental genes as maternal, zygotic and maternal-zygotic by determining which transcripts are already present in the egg and which ones are not. Maternal genes are defined as genes which mRNA is shed within the egg by the mother and are never transcribed by the embryo. Thus, the embryo contains mRNAs coming from two different genomes, the one of the mother and the one of the embryo. Maternal-zygotic genes are genes which mRNA is shed in the egg by the mother but are also transcribed by the embryo. Zygotic genes are genes which mRNA is not shed in the egg by the mother but transcribed by the embryo itself. To compare the selective regimes of the three categories a permutation test was applied (see Methods for details, section 2.3.1). No significant differences between maternal, maternal-zygotic and zygotic genes were found for ω_a (Figure 3.12). This implies that the large ω_a of the earliest stages is not due to any specific gene category. Consistent with the hypothesis of lower efficiency of natural selection on maternal genes both ω (p-value = 0.024) and ω_{na} (p-value = 0.003) were higher for maternal genes than for zygotic genes (and intermediate for the maternal-zygotic genes). On the contrary, zygotic genes show lower values than expected in the permutation test for ω_a (p-value = 0.035) and ω_{na} (p-value = 0.036). Table B.23 contains permutation p-values.

Finally, this analysis was repeated but using the genes that are in common with the genes expressed in the first fours hours according to modENCODE, to check whether maternal genes account for the high ω_{na} observed in Figure 3.9). Results were very similar, indicating that the high ω_{na} values in the earliest stages are probably due to the maternal genes in these stages (Figure A.16).

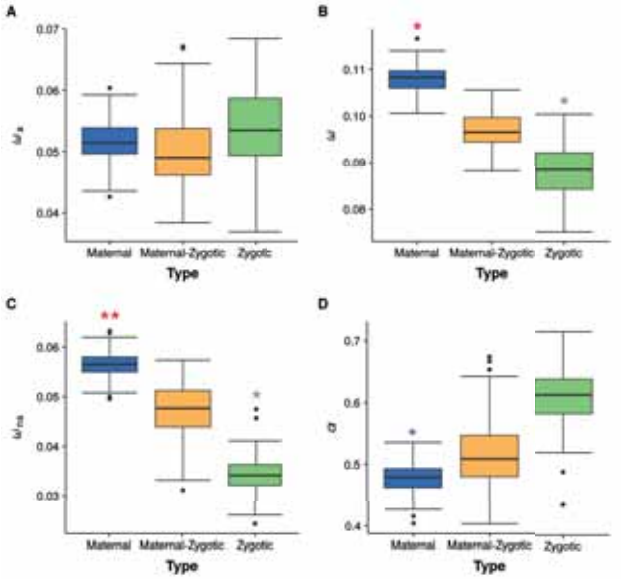


Figure 3.12 Selective regimes (ω_a, ω, ω_{na} and α) for maternal, maternal-zygotic and zygotic genes. Maternal genes are those genes which mRNA is shed by the mother in the egg and are never zygotically transcribed, maternal-zygotic are those genes which mRNA is present in the egg but that are also transcribed by the zygote. Zygotic genes are genes which mRNA are not shed in the egg by the mother. **A**. ω_a is not statistically different between these gene categories. **B**. ω is significantly higher in maternal genes than in the other two gene categories (permutation test, *p*-value = 0.024). **C**. ω_{na} is significantly higher in maternal genes than in the other two gene categories (permutation test, *p*-value = 0.003). **D**. α is marginally lower in maternal genes compared to the other two categories. Each point in a plot (100 bootstrap replicates per group) is calculated for a randomly drawn sample of the set of genes in each gene category. The number of genes analyzed in each category is shown in Table 2.4. *P*-values in table B.23.

3.3.5. Adaptation over the whole embryo's anatomy

In this last analysis, we aim to measure adaptation and constraint in the whole embryo anatomy. The integration of transcriptomics and population genomics can be used to understand the genetic and developmental basis of the phenotypic change.

From the BDGP database (Tomancak et al., 2007), genes that are expressed in 18 different anatomical organs during six different embryo developmental stages were collected. Then, this data was integrated with population genomic data to infer adaptation and constraint in different body parts.

3.3.6. A novel permutation test approach

From a statistical point of view, analyzing this anatomical data is challenging. First, a different number of genes is expressed in each anatomical term. Second, a proportion of these genes are shared between anatomical terms and/or stages. The anatomical structures are not independent and therefore a statistical test taking into account this dependency should be applied. Methods correcting for multiple testing problems (e.g., Bonferroni) are too conservative and the interpretation of a finding depends on the number of performed tests.

We have developed a novel permutation test approach that implies the advantages of the classical permutation test procedure and, additionally, overcomes the problem of multiple testing. The main advantages of permutation tests are that they can be applied to any statistic and the generated null distribution is empirical, i.e., is obtained by using the observed data (Berry, Mielke, and Johnston, 2016). Our innovation for analyzing this anatomical data consists on generating the expected null distribution simultaneously for all anatomical terms. As a result, in each permuted dataset, the number of genes co-expressed between anatomical terms stays as in the original data. See Methods, section 2.3.2, for details.

3.3.7. Action of natural selection at the germ layer level

Differences in the selective regimes experienced by the tissues derived from each of the three primary germ layers of the *D. melanogaster* embryo were assessed. These germ layers constitute the first three tissues in embryonic development: ectoderm, mesoderm, and endoderm. Later embryonic and larval tissues develop from one of the three germ layers. The set of genes that are exclusively expressed in the derivatives of each germ layer was analyzed. Therefore, genes which expression overlapped for two or three layers were excluded from the analysis. The number of genes analyzed for each germ layer is provided in Table 2.6.

The set of genes exclusively expressed in the ectoderm-derived tissues are more constrained than those expressed in the other two layers (low ω, permutation test, p-value = 0.004). On the other hand, the set of genes expressed in the tissues derived from the mesoderm show higher rates of adaptive substitutions (high ω_a, permutation test, p-value < 0.001). Finally, the set of genes expressed exclusively in the tissues derived from the endoderm show a relative relaxation of selection compared to the other two layers (high ω_{na}, permutation test, p-value = 0.046).

Neither mutation, recombination nor gene density rates differ between the genes expressed in each germ layer. Hence, these genome variables do not seem to bias our measurements of differential selection (analysis of variance, Tables B.24–B.26).

3.3.8. Selection at the anatomical structure level

The set of genes expressed in 18 anatomical structure reported in the BDGP (Tomancak et al., 2007) was analyzed. A gene was counted as expressed in a given anatomical structure if it was expressed in at least one developmental stage of the structure. The studied anatomical structures were "Amnioserosa/Yolk", "Procephalic Ectoderm/CNS", "Peripheral Nervous System (PNS)", "Foregut", "Ectoderm/Epidermis", "Tracheal System", "Salivary Gland", "Hindgut/Malpighian tubules", "Mesoderm/Muscle", "Head Mesoderm/Circulatory", System/Fat body", "Garland cells/Plasmatocytes/Ring gland", "Germ line", and "Endoderm/Midgut". In addition, genes that are expressed ubiquitously or that are present already in the egg were also analyzed. These latter genes were categorized either as "Ubiquitous" or "Maternal" according to the

original BDGP database (Tomancak et al., 2007). The number of genes per anatomical structure can be found in Table 2.5. A visual representation can be seen in Figure 2.8.

All four analyzed selective regimes –ω, ω_a, ω_{na} and α– vary through the embryo anatomy (Figure 3.13). The genes expressed in the anatomical structure "Garland cells/Plasmatocytes/Ring gland" and these expressed in the "Germ line" exhibit high rates of adaptive substitution (higher than the expected rate in random permutations of the genes in the database: high ω_a, permutation test, p-value = 0.018 and high ω_a, permutation test, p-value = 0.018, respectively). The same was found for those genes expressed in the "Head mesoderm/Circulatory system/Fat body", but only with a marginal significance (high ω_a, permutation test, p-value = 0.052).

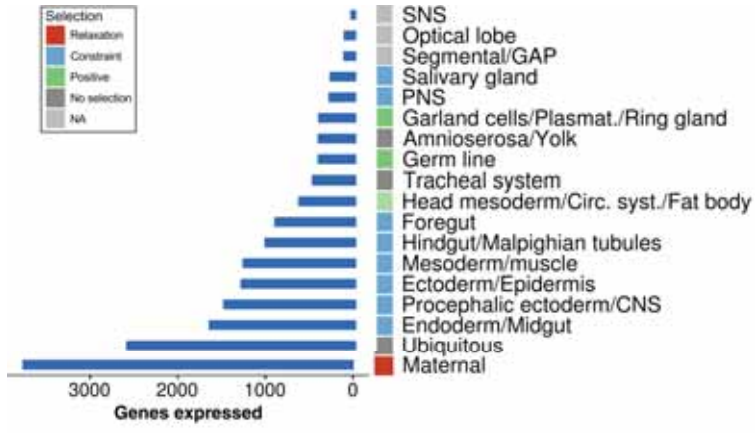


Figure 3.13 Number of analyzed genes for each anatomical term and evidence of selection. Red: relaxation of selection (high fixation of non-adaptive substitutions); blue: selective constraint (low fixation of non-synonymous substitutions); green: positive selection (high fixation of adaptive substitutions); dark grey: no evidence of selection; light grey: not analyzed, because the minimum number of 150 genes was not reached.

Contrastingly, several anatomical structures of the digestive system exhibit a high constraint in the genes they express (higher than expected from the permutation test). This is the case of the "Foregut" (low ω, permutation test, p-value < 0.001, low ω_{na}, p-value = 0.012), the "Hindgut/-Malpighian tubules" (low ω, permutation test, p-value < 0.001, low ω_{na}, p-value = 0.018), the "Endoderm/Midgut" (low ω, permutation test, p-value < 0.001, low ω_{na}, p-value = 0.018, low ω_a, p-value = 0.024) and the "Salivary gland" (low ω, permutation test, p-value < 0.001). In several

neuroectodermic anatomical structures, the set of genes expressed also showed higher selective constraint than expected by chance alone. This is the case of the peripheral nervous system (PNS) (low ω, permutation test, p-value < 0.001, low ω_a, p-value = 0.016) and the "Procephalic Ectoderm/CNS" (low ω, permutation test, p-value = 0.004, low ω_{na}, p-value < 0.001). Higher constraint was also found in the "Ectoderm/Epidermis" (low ω, permutation test, p-value < 0.001, low ω_{na}, p-value = 0.030, low ω_a, p-value = 0.024) and the "Mesoderm/Muscle" (low ω, permutation test, p-value = 0.016).

Finally, the "Maternal" genes category exhibit higher values of relaxed selection (high ω_{na}, permutation test, p-value = 0.026), as it was found in the previous study (section 3.3.4). The set of genes in the anatomical structures "Ubiquitous" and "Amnioserosa" genes do not show evidence of any preferential regime of selection.

Very similar results were found when short introns, instead of 4-fold degenerated sites, were used to estimate the mutation rate (Table B.27). Neither recombination rates nor gene density or mutation rates differ between the genes expressed in each anatomical structure (analysis of variance, Tables B.28-B.30).

3.3.9. Analysis by embryo developmental stages

The previous anatomical structures were further analyzed by splitting them between stages. In other words, each set of genes expressed in an anatomical structure and stage were analyzed independently (even for the genes expressed in the same anatomical structure at some other stage). A total of six developmental stages that span the first 16 hours of embryo development were analyzed: stage 1 (1–3), stage 2 (4–6), stage 3 (7–8), stage 4 (9–10), stage 5 (11–12), and stage 6 (13–16). The list of genes analyzed by anatomical structure and developmental stage can be found in Table B.10. Figure 3.14 shows the results obtained in this analysis. Table B.31, shows the p-values of the permutation tests for each anatomical structure, and Figure 3.15, shows a schematic illustration of the results.

In general, the results are very similar to the ones in the previous section. First of all, evidence of relaxation is only found in the first stage, where maternal genes are expressed. The following stages mainly ex-

hibit evidence of selective constraint in most structures. The anatomical structures in stage 13–16 (an embryonic stage close to the larval stage) are the ones that most often exhibit ω_a and ω values which are significantly different from the ones expected by chance.

Very similar results were found for most of the anatomical structures when short introns were used as a proxy to estimate the neutral mutation rate (Table B.32).

Figure 3.14 Anatomical structures under preponderant selection for six embryo developmental stages. **Stage 1**: Relaxation on "Maternal". **Stage 2**: Selective constraint on "Ectoderm/Epidermis" and "Procephalic ectoderm/CNS" and positive selection on "Germ line." **Stage 3**: Selective constraint on the Intestinal tract ("Hindgut/-Malpighian tubules" and "Endoderm/Midgut") and positive selection on "Germ line". **Stage 4**: Selective constraint on "Mesoderm/Muscle" and on the Intestinal tract ("Hindgut/Malpighian tubules" and "Endoderm/Midgut"). **Stage 5**: Selective constraint on Intestinal tract ("Hindgut/Malpighian tubules", "Foregut" and "Endoderm/Midgut"), "Procephalic ectoderm/CNS" and "Tracheal system". **Stage 6**: Selective constraint on "PNS", "Procephalic ectoderm/CNS", "Ectoderm/Epidermis", Intestinal tract ("Hindgut/Malpighian tubules", "Foregut", "Salivary glands" and "Endoderm/Midgut"), and positive selection on "Head mesoderm/Circulatory system/-Fat body" and "Germ line". Not shown: Stage 3: Selective constraint on "Ubiquitous" and "Ectoderm/Epidermis". Stage 4: Selective constraint on "Ubiquitous" and "Ectoderm/Epidermis". Stage 5: Selective constraint on "Ubiquitous", "Ectoderm/Epidermis", "Head mesoderm/Circulatory system/Fat Body". **Stage 6**: Selective constraint on "Ubiquitous", "Mesoderm/Muscle" and positive selection on "Garland/Plasmatocytes/Ring gland." See text for *p*-values and Figure 3.15 for a schematic version of this figure. Because several anatomical structures under constraint overlap in the figure, some are represented in dark blue and some in light blue to facilitate visualization. Abbreviations: *amg*, anterior midgut rudiment; *pc*, pole cells; *hg*, hindgut; *pmg*, posterior midgut rudiment; *hms*, head mesoderm; *ms*, mesoderm; *mp*, Malpighian tubules; *fb*, fat body; *mg*, midgut; *go*, gonads; *sg*, salivary glands. Images modified from Hartenstein (1993) with permission.

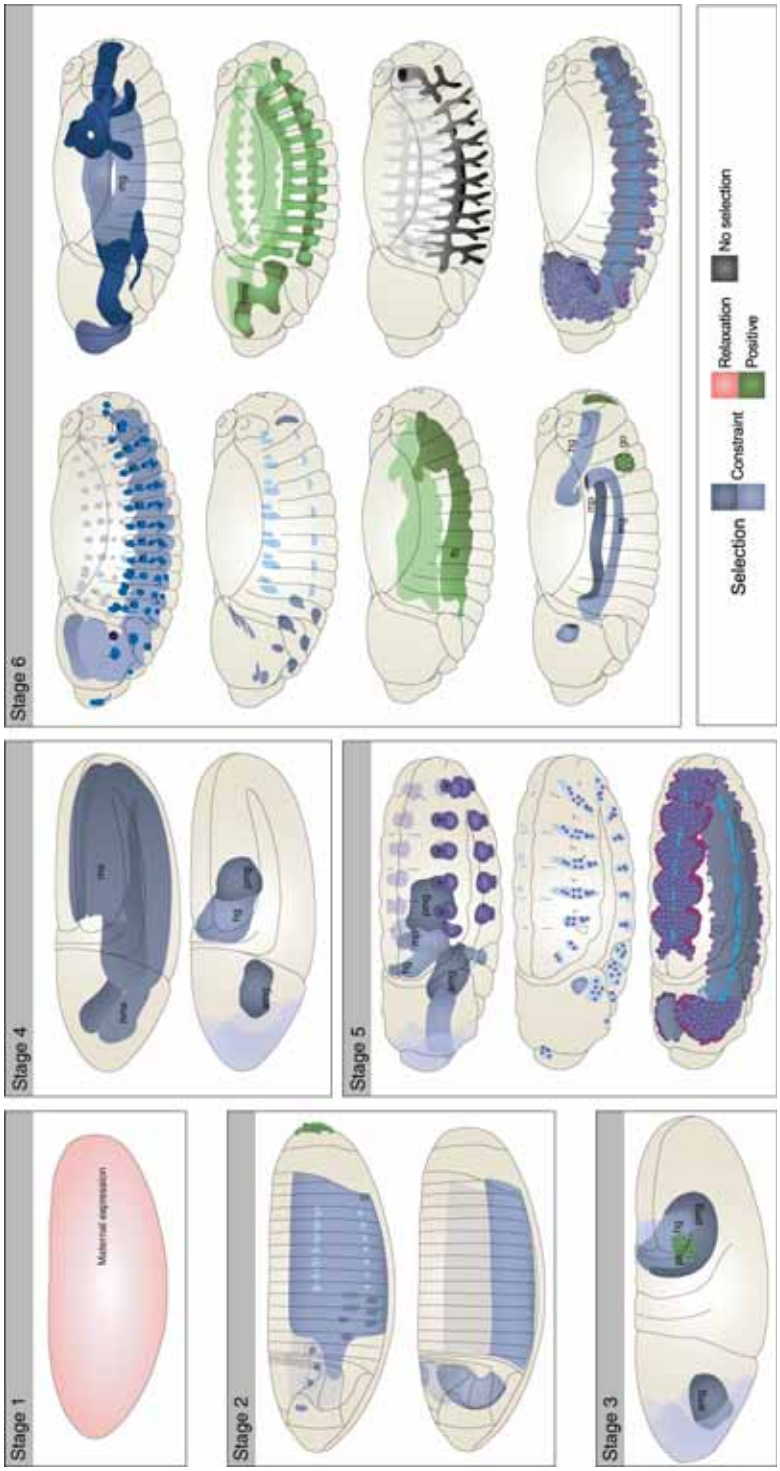


Figure 3.14 Preponderant selection on the genes expressed in each anatomical structure among stages. Caption in next page.

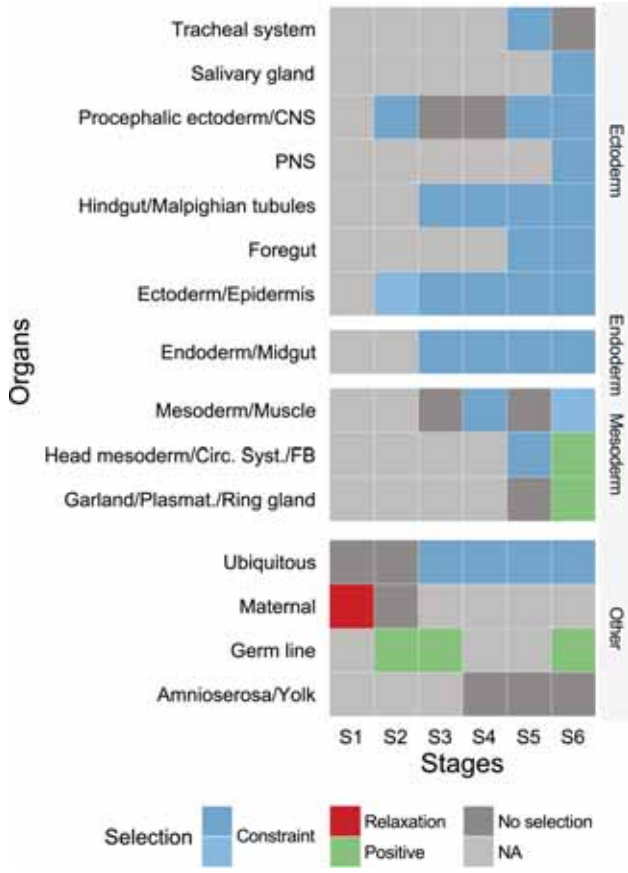


Figure 3.15 Summary of evidence of selection on the genes expressed in each anatomical structure among stages. Schematic version of Figure 3.14. Color patterning as in Figure 3.13.

3.3.10. Relationship between phylogenetic age, *Fop*, expression bias and adaptation

The relationship between the phylogenetic age (using Drost, 2014 data), the expression bias and expression level (using modENCODE RNA-seq expression data, Graveley et al., 2011), the frequency of optimum codons (*Fop*), and the different selection regimes was analyzed. Additionally, the set of genes expressed in the 18 anatomical structures were divided into eight groups depending on the number of anatomical structures in which they are expressed (1, 2,..., 7, 8 or more). These values can be

taken as a rough measurement of the pleiotropic effects of a gene on embryonic anatomy. This index is called *spatial pleiotropy*.

As shown in Figure 3.16, the anatomical structures with the highest rates of adaptive substitutions are not the anatomical structures with the lowest *Fop*, newest genes, or highest expression bias as it was expected according to the patterns found in section 3.2. Therefore, these variables do not seem to explain the differences in the rates of adaptive substitution found between anatomical structures.

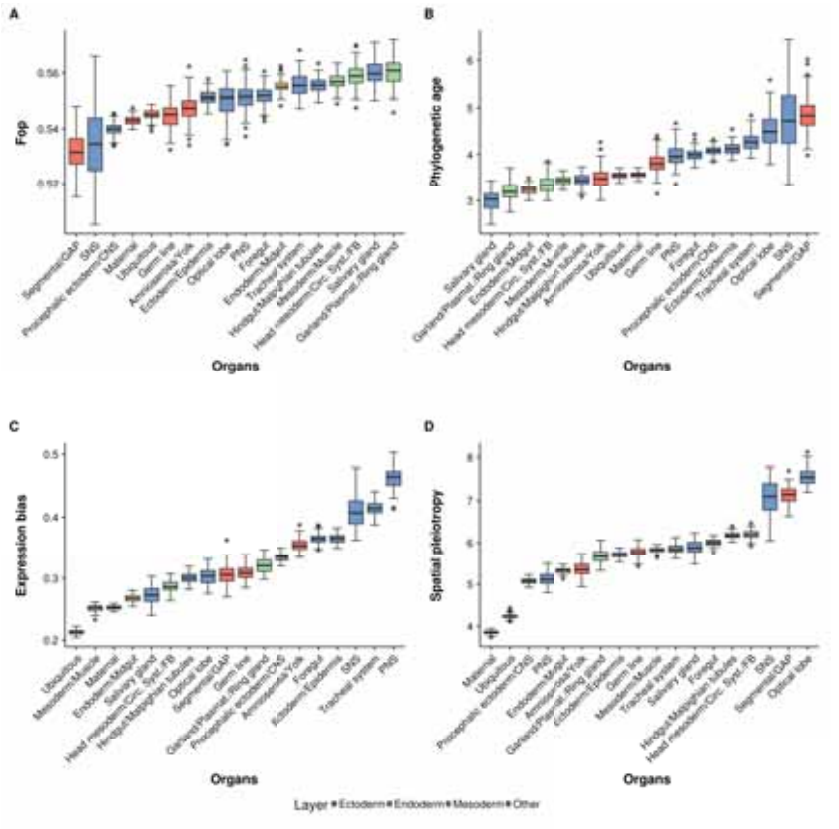


Figure 3.16 Mean resampling of *Fop*, phylogenetic age, expression bias and spatial pleiotropy in each anatomical structure. **A**. *Fop* mean resampling. **B**. Phylogenetic age mean resampling. **C**. Expression bias mean resampling. **D**. Spatial pleiotropy mean resampling. Resampling of each anatomical structure is estimated by resampling with replacement 100 times the genes in each anatomical structure.

3.3.11. Relationship between phylogenetic age, *Fop* and expression bias

To acquire a better understanding of the results in the previous section, the relationship between the phylogenetic age, expression bias, and *Fop* of the analyzed genes was assessed.

A positive correlation between phylogenetic age and expression bias was found (Pearson's ρ = 0.49, *p*-value = 0.039, Figure 3.17). Thus, younger genes tend to be expressed in more specific stages than phylogenetically older genes, which are more broadly expressed through stages. Furthermore, genes expressed in anatomical structures derived from the endoderm are phylogenetically the oldest on average, whereas those derived from the ectoderm express the youngest genes (except for the set of genes expressed in the salivary glands).

The "Segmental/GAP" anatomical structure is also exceptional in expressing the youngest genes (note that during development these genes are expressed before the germ layers are formed). A negative correlation is found between phylogenetic age and *Fop* (Pearson's ρ = -0.698, *p*-value = 1.27×10^{-3}). The salivary glands stand out for having one of the highest *Fop* values.

3.3.12. Relationship between pleiotropy, phylogenetic age, *Fop*, expression bias and adaptation

The relationship between the spatial pleiotropy and the phylogenetic age, *Fop*, and expression bias was analyzed.

A negative correlation between the spatial pleiotropy and the phylogenetic age was found (Figure 3.18A, Pearson's ρ = 0.777, *p*-value = 0.023). A negative correlation was also found for expression bias (Figure 3.18B, Pearson's ρ = -0.9 *p*-value = 0.002). Finally, a positive correlation was found with *Fop* (Figure 3.18C, Pearson's ρ = 0.926, *p*-value = 9.51×10^{-4})

When the patterns of selective regimes were analyzed, a negative correlation between spatial pleiotropy and both ω (Figure 3.19A, Pearson's ρ = -0.89, *p*-value = 0.003) and ω_{na} (Figure 3.19B, Pearson's ρ = -0.749, *p*-value = 0.032) was found.

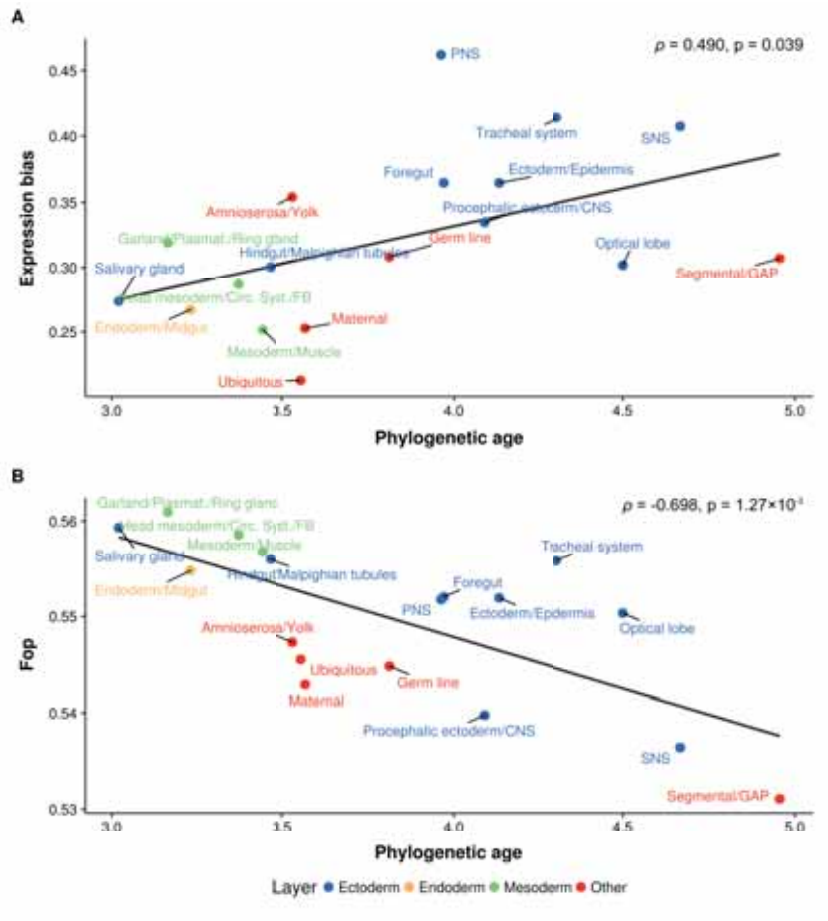


Figure 3.17 Relationship between phylogenetic age and expression bias and *Fop*. **A**. Positive correlation between phylogenetic age and expression bias. **B**. Negative correlation between phylogenetic age and *Fop*. Each dot represents the mean of each anatomical structure. Each color represents the germ layer of origin of each anatomical structure: blue: ectoderm origin; yellow: endoderm origin; green: mesoderm origin; red: not originated from the germ layers.

Thus, genes expressed in a low number of anatomical structures seem to be less selectively constrained than genes expressed in a high number of anatomical structures. No correlation was found between ω_a and the number of anatomical structures in which a gene is expressed (Figure 3.19C). As shown in Figure 3.16D, the anatomical structures with the highest ω_a are not the ones where lowest spatial pleiotropy. Therefore, these results are not simply explainable from differences in *Fop*, phy-

logenetic age, spatial pleiotropy or expression bias among anatomical structures.

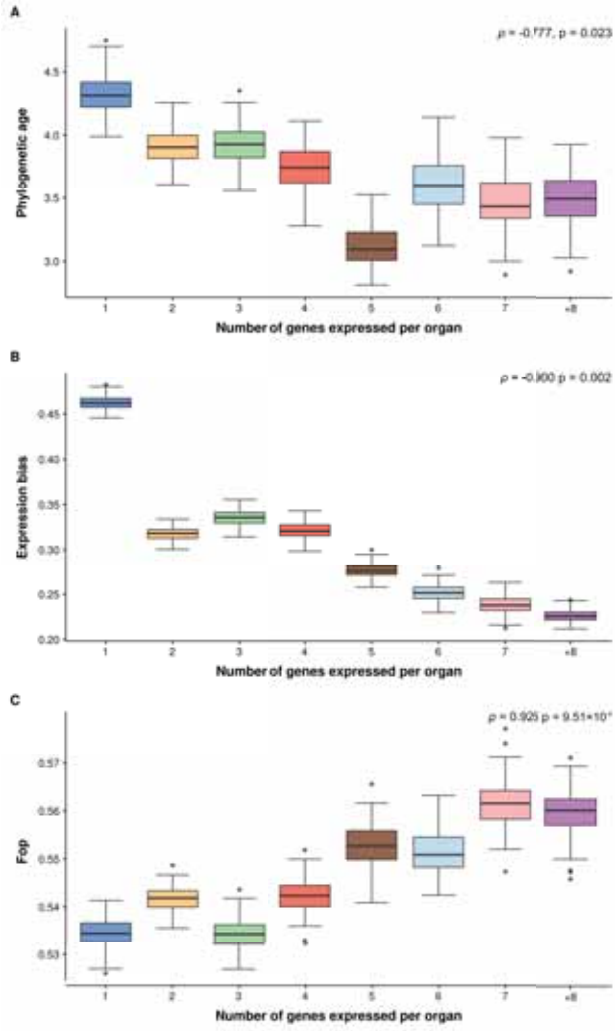


Figure 3.18 Relationship between spatial pleiotropy, phylogenetic age, expression bias and *Fop*. The gene data set was divided into eight groups depending on the number of anatomical structures in which they are expressed (1, 2, ..., 7, 8 or more). Each group is obtained by resampling 100 times with replacement the genes of each group. **A**. Negative correlation between the spatial pleiotropy and phylogenetic age. **B**. Negative correlation between spatial pleiotropy and expression bias. **C**. Positive correlation between the spatial pleiotropy and the *Fop*.

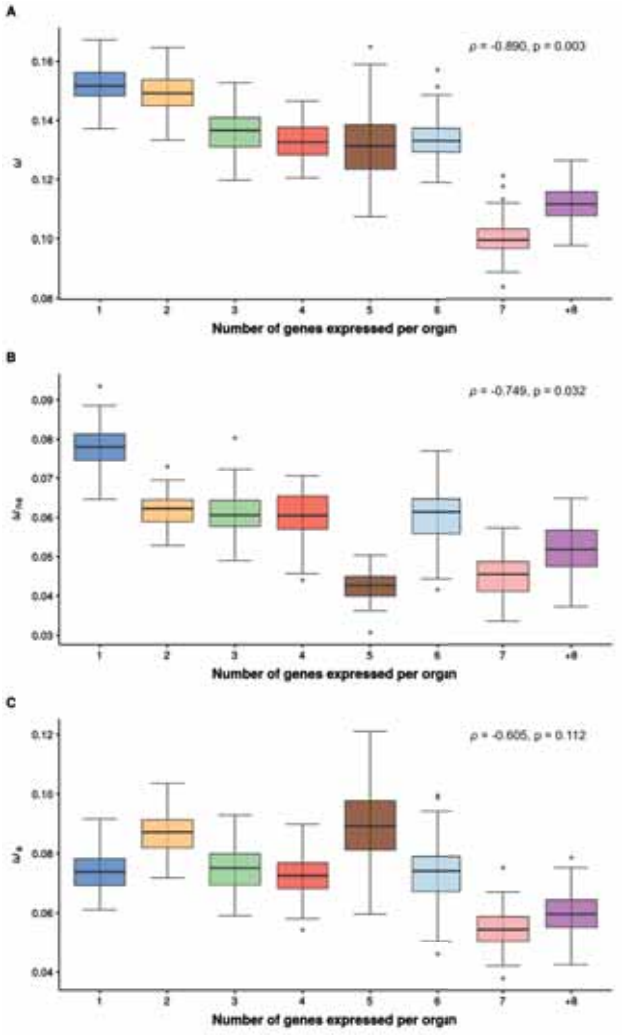


Figure 3.19 Relationship between spatial pleiotropy and ω, ω_{na} and ω_a. **A**. A negative correlation between ω and spatial pleiotropy is found. **B**. A negative correlation between ω_{na} and spatial pleiotropy is found. **C**. No correlation between ω_a and the gene groups is found. Each group is estimated by resampling 100 times with replacement the genes in each group.

Chapter 4

DISCUSSION

4

Discussion

Molecular population genetics was born half a century ago. During these years, great progress and changes in data acquisition and theoretical developments have revolutionized the field (Casillas and Barbadilla, 2017). Today, this genomic revolution has provided us with enough detailed population genetic data, which in combination with sophisticated statistical methodologies, allows us the large-scale analysis of genomic patterns of DNA variation (Casillas and Barbadilla, 2017).

This thesis represents a complete analysis of the three population genomics levels in a species (Figure 4.1): the DNA variation level, the genomic level and the multiomic (integrative) level (Casillas and Barbadilla, 2017). These three levels are consecutive and interconnected since every level acts as the input for the subsequent one. The first part of the thesis consisted of the evaluation of a set of MKT statistics, which required first the compilation of the necessary population genomics data and the estimation and description of selection parameters. In the second part, the previous estimates were correlated with features estimated along the genome to assess their relative importance on the molecular

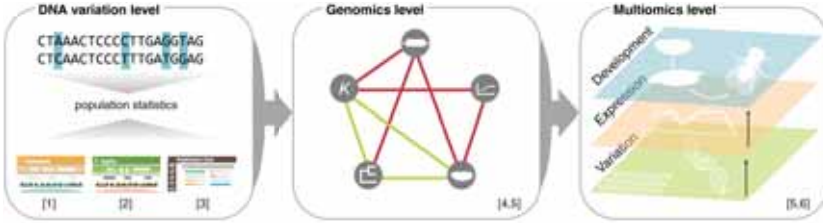


Figure 4.1 The inquiry power of population genomics approach. Representation of three population genomics level, emphasizing the contributions provided by the BGD group and the present thesis. Contributions [1] Casillas et al. (2018); [2] Hervas et al. (2017); [3] Murga-Moreno et al. (2018); [4] Castellano et al. (2015); [5] Coronado-Zamora et al. (submitted); [6] Salvador-Martínez et al. (2018).

evolutionary rate of protein-coding genes. Ultimately, in the third part, the patterns of genomic diversity were integrated with other multiple -omics layers across developmental time and space. On the next pages, the main results and impact of the findings are discussed.

4.1. Population genomics at the DNA variation level

The availability of molecular techniques and bioinformatic tools that makes the study of genome variation within and between species possible has undoubtedly been one of the greatest advances in this genomic revolution (Charlesworth and Charlesworth, 2017). Nowadays, NGS technologies allow the fast and cheap sequencing of thousands of genomes. This huge amount of data, together with sophisticated methods to analyze it, can reveal the meaning of the variability existing at the DNA sequence level.

One of the most comprehensive and complete population genomics community resources, the *Drosophila* Genetic Reference Panel (DGRP), has been used for testing molecular population genetics hypotheses and estimating the fraction of adaptive evolution, α. During the last years, several statistical methods have been developed for quantifying the amount of selection in a genome using polymorphism and divergence data. The McDonald and Kreitman test (McDonald and Kreitman, 1991) and its derivatives are extensively used to detect the signature of natural selection at the molecular level. Their main advantages and more importantly, their limitations, are discussed in the following section 4.1.1.

4.1.1. Estimating the adaptive rate in *D. melanogaster* with MKT-based methods

The first analysis of the thesis consisted in the general description of the patterns of polymorphism and divergence in a total of 13,753 protein-coding genes of a North American population of *D. melanogaster*. 76.38% of the genes exhibit variation both in polymorphism and in divergence, an indispensable requirement for applying the MKT and for quantifying the proportion of adaptive substitutions (α).

There are four important factors that must be taken into account when performing an MKT, as they can imply a rejection of the neutral hypothesis and significantly affect the estimates of adaptive evolution. Those factors are: (i) the outgroup for estimating divergence parameters, (ii) the effect of the segregation of slightly deleterious substitutions, (iii) the constancy of the neutral mutation rate over time and (iv) the effect of concatenating sites. The impact of these four factors in the presented results are thoroughly discussed in the next sections.

Impact of the chosen outgroup on the divergence estimates

The estimates of divergence were quantified in two close sister species of *D. melanogaster*, *D. simulans* and *D. yakuba*. *D. melanogaster* and *D. simulans* diverged approximately 4.3 Mya (Cutter, 2008), whereas the divergence time with *D. yakuba* is longer, 7.4 Mya (Tamura, Subramanian, and Kumar, 2004). Keightley and Eyre-Walker (2012) determined that estimating the rate of adaptive evolution can be biased especially when the divergence time between two species is low relative to within-species variation. As Keightley and Eyre-Walker (2012) summarize, this bias can be due to (i) an erroneous attribution of polymorphism to divergence (Figure 4.2), (ii) ancestral polymorphism contributing to divergence (Figure 4.2) and (iii) differences in the rate of fixation of neutral and adaptive mutations. The first bias can happen because, typically, divergence is calculated by randomly selecting an allele from the focal species and comparing it to the allele in the outgroup species (for which generally just one sequence is available). However, some differences categorized as divergent could be, in reality, due to polymorphisms. The problem can be partially addressed taking into account all available alleles, as it is unlikely that a polymorphism will appear as fixed in a large sample of sequences. Inflating divergence by this polymorphism misattribution can lead to an overestimation of the rate of adaptive evolution, α (Keightley and Eyre-Walker, 2012). An example of the second case should be the segregation of a slightly deleterious substitution. It is more likely that a slightly deleterious substitution segregating at the time of the divergence will be lost in one lineage and continue segregating in the other one. As a consequence, ancestral polymorphism originating from slightly deleterious substitutions contribute more to divergence than neutral mutations, leading to an overestimation of α (Keightley and Eyre-Walker, 2012). Again, this bias can be partially solved taking into account all possible alleles. However, as noted above, polymorphism

data is usually only available for the focal species and not for the outgroup. Furthermore, unlike the minor allele count that can be directly observed from sequence polymorphism data, the inference of the ancestral state requires maximum parsimony methods that can potentially produce misleading results Keightley and Jackson, 2018. In the third scenario, advantageous mutations reach fixation faster than neutral ones, inflating the adaptive estimates in the short term (Keightley and Eyre-Walker, 2012) –this scenario is further argued when the neutral mutation rate is discussed below.

The fact that a neutrality index (NI) of 0.77 was obtained when using *D. simulans* as outgroup and an NI of 1.06 when using *D. yakuba*, can be explained in three ways. First, *D. simulans* has experienced more adaptive evolution (i.e. D_N is higher than expected), which is a rather straightforward explanation. Second, the estimation is affected by the short divergence time. According to Bierne and Eyre-Walker (2004), because the effective population size of *D. simulans* is higher than that of *D. melanogaster*, the former had less time of fixation (the coalescence time is on average longer). Therefore, neutral mutations will have less time to fix in the *simulans* lineage. In contrast to neutral mutations, beneficial mutations spread much more rapidly through a population than neutral do. Therefore, using *D. simulans* will likely lead to inflated α estimations (Bierne and Eyre-Walker, 2004). However, this reason seems to contradict the average negative α that was obtained when the MKT was performed in individual protein-coding genes (Table 3.1). Then, the third and more plausible explanation is that the gene heterogeneity is high in the *D. melanogaster-D. simulans* gene dataset. That is, there are large differences in the number of non-synonymous substitutions (D_N) between genes, leading to the Simpson's paradox (see section 1.1.4). In fact, using the weighted NI statistic yields a NI_{TG} of 0.98, a more reasonable value according to the average α.

The recent availability of genomic data of a North American *D. simulans* population (Signor, New, and Nuzhdin, 2018) represents a valuable complement to the DGRP and other *D. melanogaster* panels. Such resources can reduce the bias associated with polymorphisms contributing to apparent divergence, although polymorphism may still appear to be fixed in a sample of sequences (Keightley and Eyre-Walker, 2012).

Tataru et al. (2017) recently presented an approach that proposes a hierarchical probabilistic method to infer α from polymorphism data alone

without using divergence data. This is a promising complement to obtain better estimates of adaptive molecular evolution independently of an outgroup. On the opposite, Keightley and Jackson (2018) proposed a maximum-likelihood method to estimate the SFS of a focal species using information from multiple outgroups (up to three) while assuming simple models of nucleotide substitution.

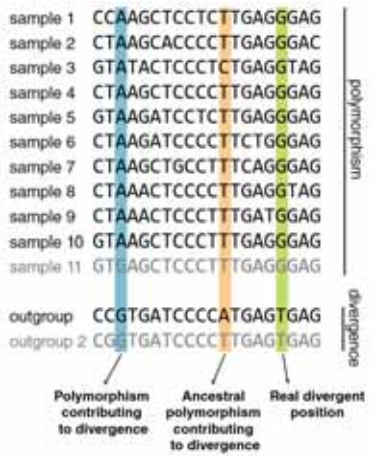


Figure 4.2 Representation of two possible misatributions of polymorphism to divergence. The sequences used for the computation of divergence are represented in black. In gray, the ones that are not used and provide additional information on the alleles status.

Segregation of slightly deleterious substitutions

Another important factor affecting the estimates of adaptive evolution is the segregation of slightly deleterious substitutions, as it has been already introduced in section 1.1.4. The impact of slightly deleterious substitutions has been assessed comparing five MKT approaches. Slightly deleterious substitutions violate the assumption that only neutral mutations contribute to polymorphism in an MKT. Several methods (Templeton, 1996; Akashi, 1999; Fay, Wyckoff, and Wu, 2001; Bustamante et al., 2002; Smith and Eyre-Walker, 2002; Sawyer et al., 2003; Bierne and Eyre-Walker, 2004; Bustamante et al., 2005; Mackay et al., 2012; Messer and Petrov, 2013) have attempted to circumvent this effect by excluding or taking into account the polymorphism at low frequencies, because slightly deleterious substitutions segregate at a low level. The MKT-based methods that have been compared in the present thesis are the standard MKT McDonald and Kreitman, 1991, the FWW method (Fay, Wyckoff, and Wu, 2001), the eMKT (Mackay et al., 2012), the asymptotic MK (Messer and Petrov, 2013) and a derived method developed by the BGD group, iMKT. Within the MKT-based methods, the DFE-alpha

approach was not included. The reason was that is a sophisticated likelihood approach that also models demography, in contrast to the previous ones. However, as DFE-alpha is widely used for the estimation of the proportion of adaptive substitutions, is discussed in section 4.1.2.

The comparative analysis of the MKT-derived methods allows us to conclude the following: first, the assumptions of the standard MKT do not hold for *D. melanogaster* data. As Bierne and Eyre-Walker (2004) pointed out, the use of *D. melanogaster* data is likely underestimating α, unless a methodology correcting for the segregation of slightly deleterious mutations is used. The FWW correction (Fay, Wyckoff, and Wu, 2001) exclude polymorphisms in both the neutral and the selected site classes (P_S, P_N) at frequencies below a given threshold. However, this method is still expected to lead to biased estimates because slightly deleterious substitutions can still segregate at a frequency above the threshold (Charlesworth and Eyre-Walker, 2008), and removing data implies the loss of statistical power. The eMKT proposed by Mackay et al. (2012), instead of removing the polymorphism at low frequencies, separates P_{N}, the count of segregating sites in the non-synonymous class, into the number of neutral variants and the number of weakly deleterious variants. This allows the evaluation of adaptive and weakly deleterious independently while increasing the statistical power to detect selection. The asymptotic method proposed by Messer and Petrov (2013) is conceptually the best because it allows an efficient removal of the slightly deleterious substitutions in all frequencies and not below a conventional threshold as in the FWW and eMKT methods. On the other hand, this procedure lacks power when applied to individual genes. Even though the number of differences between *D. melanogaster* and its outgroup species and within *D. melanogaster* is high, it is not enough for the iMKT to work for single genes. For this reason, such methodologies use concatenated sets of genes to estimate the value of α overall (Boyko et al., 2008; Eyre-Walker and Keightley, 2009). The gene concatenation process is discussed in detail in the following section.

Simulations via SLiM 2 has been conducted as a benchmark to compare the performance of the mentioned methodologies under different evolutionary scenarios. One important advantage of performing simulations is that a predefined α value is known, and therefore the methodology estimating an α value closest to this known value can be assessed. iMKT is the best method in terms of estimating the most accurate α value, but its performance decreases in simulations which generated less poly-

morphism (a shorter genomic region, less generational time or lower mutation rate). In those cases, iMKT performance was similar to FWW and eMKT corrections.

Haller and Messer (2017) advise using a cutoff of [0.1,0.9], thus removing polymorphism at a frequency lower than 0.1, implicitly applying an FWW-like correction. If the estimations are repeated without applying this cutoff, there is almost no difference between the α values estimated by iMKT and FWW (Figure 4.3), especially in those scenarios that produced less polymorphism. And, on top of that, the iMKT performance diminishes in such scenarios, especially when the simulated chromosome is short (only 90% of the simulations runs could be analyzed by iMKT) because the polymorphism simulated is not enough.

Constancy of the neutral mutation rate over time

One of the most unrealistic assumptions of the MKT is that the neutral mutation rate is constant over time (Hahn, 2018), meaning that the selective constraint is also constant. However, the neutral rate is heavily affected by changes in the effective population size (N_e) (Balloux and Lehmann, 2012; Lanfear, Kokko, and Eyre-Walker, 2014).

For an illustrative example of why neither neutral mutation rate nor selective constraint are constant over time, one can imagine the evolutionary trajectory of duplicated genes. For a newly duplicated gene, the strength of selection is initially relaxed and then changes when a new function is acquired, becoming under selective constraint. On the other hand, for a single-copy gene, for example, the strength of selection may fluctuate over time. In these cases, the MKT results can be misleading. However, this effect is expected in general to be negligible in the MKT for single genes because the fluctuations in the selection strength over time should not have a directionality (Fay, Wyckoff, and Wu, 2001). At the population level, on the contrary, persistent changes in the population size can have a big impact on the level of sequence constraint and therefore affect the MKT considerably. For example, if a population has been expanding, slightly deleterious substitutions can lead to an overestimation of α as they could have been fixed in the past (thus, contributing to divergence) due to the larger importance of genetic drift in small populations (Eyre-Walker and Keightley, 2009).

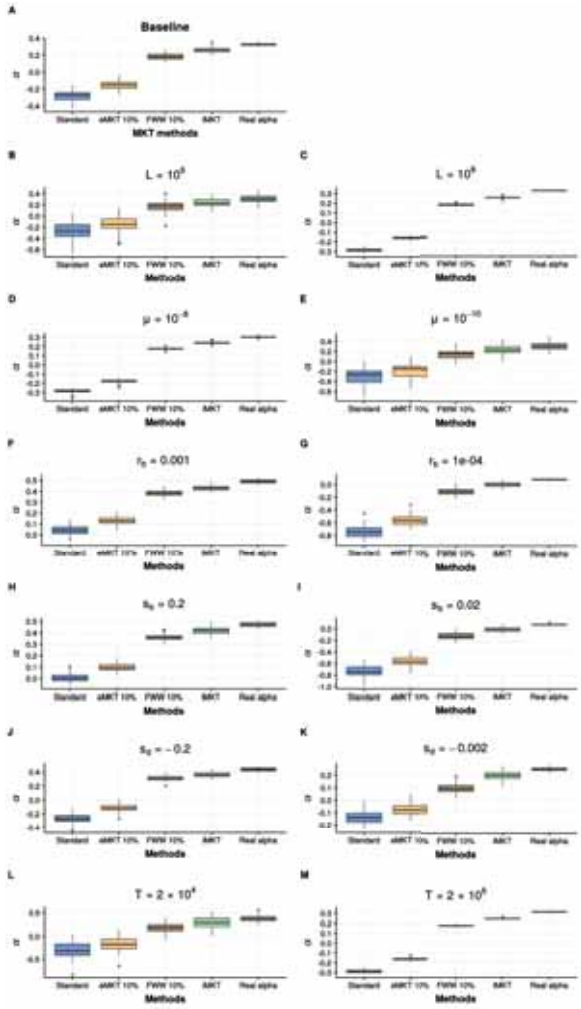


Figure 4.3 Results from the five MKT approaches for 13 simulation runs conducted with SLiM 2. A. Shows the averaged results from 50 replicate runs of the baseline SLiM model supplied on Messer & Petrov GitHub (see Methods, section 2.2.1). These runs used parameter values of mutation rate $\mu = 10^{-9}$ per base position per generation, chromosome length $L = 10^7$ base positions, beneficial mutation rate $r_b = 0.0005$, beneficial mutation selection coefficient $s_b = 0.1$, deleterious mutation selection coefficient $r_d = -0.02$, and time after burn-in $T = 2{\times}10^5$ generations. The subsequent graphs (**B-M**) shows the results from 50 replicate runs using the non-baseline parameter value shown in the graph title. A DAF of 20 frequency bins was used, with an x cutoff of [0,1]. eMKT and FWW methods corrects polymorphisms below a 10% frequency.

Proxy for the mutation rate

Due to the degeneracy of the genetic code, synonymous sites of protein-coding genes has been considered *silent* and free of natural selection. Many examples of studies using 4-fold degenerated sites are used as a proxy for the mutation rate (e.g.: Halligan et al., 2010; Halligan et al., 2013; Castellano et al., 2015; Phung, Huber, and Lohmueller, 2016). However, some studies have pointed out that these sites are subject to selection, especially due to codon usage (reviewed by Hershberg and Petrov, 2008). In *D. melanogaster*, it was found that 22% of 4-fold synonymous sites evolve under selective constraint (Lawrie et al., 2013). Regardless of the reasons why 4-fold degenerated sites could be under constraint, this has serious implications for the interpretation of the d_N/d_S test. Specifically, it can lead to an overestimation of positive selection. When 4-fold degenerate sites are constrained, the proportion of non-synonymous sites interpreted to evolve more than the neutral reference is elevated, leading to an increased rate of false positives in the detection of positively selected sites or genes (Künstner, Nabholz, and Ellegren, 2011). In the case of the MKT, because it takes into account the polymorphism, the constraint in synonymous sites would downwardly bias the α estimations.

The analyses presented here were performed using both 4-fold degenerated and short intron sites as a proxy for the neutral mutation rate, always leading to similar results. In some cases, the rate of adaptive evolution (measured with α or ω_a) was higher when 4-fold were used compared to the same estimates using short introns. *D. melanogaster* has an asymmetrical intron length distribution, with a group of short and another of long introns (Parsch et al., 2010). Short introns appear to be under less selective constraint than long introns (Parsch et al., 2010). Within short introns, the least constrained sites are those falling between the 5′ and 3′ regions of the intron which operate in splice site recognition (Halligan and Keightley, 2006). Halligan and Keightley (2006) found that the fastest evolving intron sites are the bases 8–30 of introns $\leqslant$65 bp. These findings suggest that short intron sequences may be the most appropriate proxy for the neutral mutation rate. Furthermore, Parsch et al. (2010) show that the high divergence observed in short introns is not due to adaptive evolution. There are some limitations to the use of short introns. First, Lawrie et al. (2013) noted that parts of the short introns could be under selection. Second, short intron sites are not present in all protein-coding genes, contrary to 4-fold degenerated sites. There-

fore, gene datasets can significantly differ regarding their short intron content (e.g., in the gene dataset used in this thesis, 6,690 out of 11,003 protein-coding genes have short introns).

The effect of concatenating data

As explained above, the process of concatenating genes to create single evolutionary entities is a good strategy to overcome the problem of not having enough polymorphism data to conduct an MKT. In the majority of the performed analyses, this process does not seem to affect the results. However, there are some caveats that must be taken into account when interpreting results obtained by this procedure.

First of all, genes that are in the same genomic context are not necessarily concatenated. This is of particular importance in the case of recombination because it can affect the variance of segregating sites among regions (Hahn, 2018). Concatenated genes do not necessarily share the same recombination context, GC-content or gene density rate, which are factors related to the adaptive capacity of genes Castellano et al., 2015. It is unlikely, though, that inside a gene this can have a significant effect on the MKT. But when concatenating large gene sets, the increased variance in the number of segregating sites per gene due to both the constraint of the gene *per se* and the recombination context in which genes lay (when there are no linked selection) can lead to a rejection of the neutral hypothesis or a diminished statistical power do detect selection (Hahn, 2018). As an example, the neutral site frequency spectrum (SFS) derived from a concatenating process may significantly differ from the one obtained from a specific gene, affecting the MKT (Hahn, 2018).

By concatenating hundreds of genes, it is more difficult to detect a signal of positive selection if it is only happening on a few genes of the pool. In summary, all the evolutionary forces acting differentially on different genes contribute to the dilution of potential biological signals.

4.1.2. DFE-based methods: DFE-alpha

Other methods correcting for the potential biases of the MKT estimates are based on the estimation of the DFE at functional sites (Bustamante et al., 2002; Bustamante et al., 2005; Eyre-Walker, 2006; Eyre-Walker and

Keightley, 2007; Keightley and Eyre-Walker, 2007; Boyko et al., 2008; Eyre-Walker and Keightley, 2009). The DFE-based methods first estimate how many non-synonymous substitutions are expected to get fix given the DFE and any excess on this expectation is attributed to adaptation. DFE-alpha method (Eyre-Walker and Keightley, 2009) is a widely used method belonging to the DFE-based extensions, which additionally aims to correct for possible effects of demography.

Messer and Petrov (2013) compared the performance of their method, the asymptotic MKT and the DFE-alpha method through simulations. The authors claimed that DFE-alpha correctly estimated α when the model allowed population size change, but the demography inferred was found to be wrong, mainly due to background selection acting at linked sites (Messer and Petrov, 2013). Genetic draft leaves signatures in the SFS similar to those observed under a recent population size expansion. The DFE-alpha method systematically inferred a population expansion even though no expansion was set in the simulation (Messer and Petrov, 2013).

Another limitation of DFE-alpha is that it becomes computationally intensive, especially when a two-size-change demographic model is applied (Kim, Huber, and Lohmueller, 2017). Since DFE-alpha can only consider two population-size changes, it becomes insufficient for capturing the excess of rare variants due to the complex demographic history of some populations, like the human history (Keightley and Eyre-Walker, 2007; Kim, Huber, and Lohmueller, 2017).

4.1.3. The North American population of *D. melanogaster*

D. melanogaster originated from Africa and expanded to the rest of the world (David and Capy, 1988; Lachaise et al., 1988). Around the middle of the 19th century *D. melanogaster* populations from Europe colonized North America. The *D. melanogaster* North American population used in the analyses (DGRP, Mackay et al., 2012) contains a subset of the genetic variation of the European population, which is, in turn, a subset of the African genetic variation (Caracristi and Schlötterer, 2003). However, some studies showed that the North American population is more similar to the African population than the European one (Caracristi and Schlötterer, 2003; Baudry, Viginier, and Veuille, 2004; Haddrill et al., 2005). Therefore, the North American population of *D. melanogaster* is

an admixture of populations, with a 15% of African ancestry and 85% European ancestry (Duchen et al., 2013).

A weak spot of the DGRP is the inbreeding approach followed to obtain the isolines (see Methods 2.1.1). The inbreeding approach alters the frequency spectrum of the lethal or strongly deleterious recessive mutations (García-Dorado, 2012). However, alternative resources such as the *Drosophila* Population Genomics Project (DPGP, Langley et al., 2012) would encounter the same problem. Previous works have compared adaptation and DFE estimates between DGRP and DPGP datasets, showing no differences between them (Castellano et al., 2015; Castellano, James, and Eyre-Walker, 2017). Given that DPGP sample size (n=110) is smaller than DGRP (n = 205) it is likely that these mutations contribute very marginally to the estimations of polymorphisms, DFE and adaptation in both databases. Therefore, it is expected that the DGRP isolines contain a representative sample of the natural variation of the population at the moment at which the flies were sampled (Mackay et al., 2012).

4.1.4. From *Drosophila* to humans

The availability of the most complete worldwide nucleotide variation dataset for humans, the 1000 Genomes Project (1000GP, Auton et al., 2015), provides abundant data to detect targets of positive selection in our species. It has previously been demonstrated that *Drosophila* genomes undergo pervasive positive selection (Mackay et al., 2012; Huang et al., 2014), and their high variability and high population effective size make it the perfect candidate to test adaptive evolution. In humans, on the contrary, only a few protein-coding genes have sufficient divergence and polymorphic sites to be analyzable with MKT approaches which highlight the need of other approximations to detect signals of positive selection in single-gene data, such as the Bayesian population genetic inference method (Bustamante et al., 2005).

4.2. Population genomics at the genomic level

The second step in this three population genomics level analysis is the correlation of the population genomic parameters previously estimated (α) with several properties estimated along the genome. Thanks to the

availability of more diverse genomic datasets measuring factors such as gene expression or recombination maps as well as improved genomic annotations it is possible to integrate the genomic and functional dimension with the population dimension.

α is not the only parameter that was used for performing the correlations with the patterns of genome variation. Three other closely related population statistics are also incorporated in the study: (i) ω_a, which is the proportion of adaptive substitutions relative to the mutation rate, estimated as $\omega \times \alpha$ (Gossmann, Keightley, and Eyre-Walker, 2012); (ii) ω_{na}, which is similar to ω_a, but instead of accounting for adaptive substitutions, it considers the non-adaptive ones (Galtier, 2016); (iii) ω, which is the total number of non-synonymous substitutions relative to the neutral mutation rate (the sum of ω_a and ω_{na}). ω is extensively used as a proxy for conservation at the sequence level.

An inventory of genomic features was estimated throughout the genome of *D. melanogaster*. These genomic features span from (i) gene-architectural features; (ii) expression features; (iii) genomic context features and (iv) phylogenetic features.

How the features listed above contribute to the proportion of adaptive and non-adaptive substitutions (ω_a and ω_{na}, respectively) on protein-coding genes? Previous works tried to define the main determinants of protein evolution by applying sophisticated statistical regression models to correct for the effect of other covariates in the presence of heterogeneous data (see Drummond, Raval, and Wilke, 2006; Plotkin and Fraser, 2007). Nonetheless, the pervasive evidence of multicollinearity between existent determinants of protein evolution (which can be represented performing a partial correlation matrix between all features, Figure 4.4) suggests that some genomic features cannot be split into single independent entities, and must be considered together. Next, I will discuss the most relevant relationships, some of which were already described in the literature.

The main distinction of this work in contrast to previous ones (e.g., Duret and Mouchiroud, 2000; Larracuente et al., 2008) is the estimation of α using polymorphism and divergence data by means of the DFE-alpha software.

Another distinction is the concatenation of sequences to increase the statistical power for the calculation of polymorphism and divergence

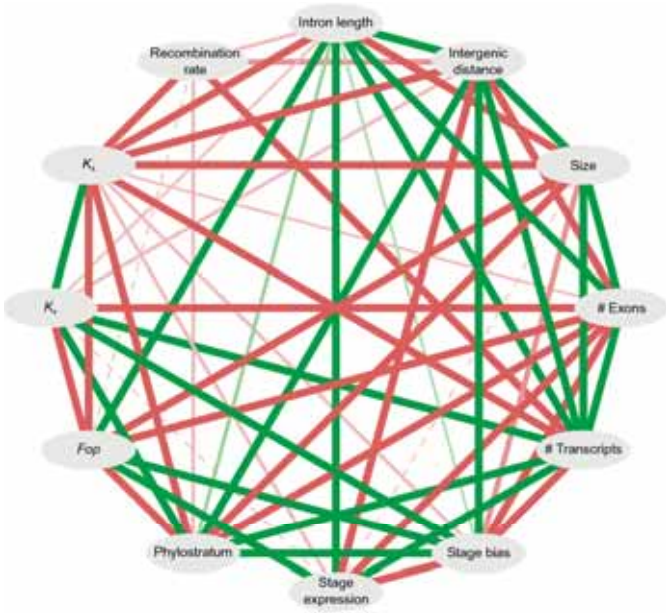


Figure 4.4 Multicollinearity of genomic features. The diagram represents all genomic features of 9,683 genes included in a partial correlation matrix. Green lines connecting two features represent positive partial correlations, and red lines represent significant negative partial correlations. The thickness of the lines corresponds to the magnitude of the *p*-value (<0.0005, <0.005 and <0.05). Dashed lines indicate that the correlation was marginally significant. Two determinants are not connected when there was no significant correlation.

metrics. As explained before, any concatenation process reduces the true (i.e., biological relevant) variation due to the heterogeneity of genes. With our approach, we tried to minimize the loss of statistical power by splitting the empirical distributions of genome features into five categories (Table B.11), maintaining an equivalent number of genes in each category. It is then possible to perform simple linear regressions between the genomic features and the estimated selective regimes.

4.2.1. Recombination and the efficacy of positive and purifying selection

The here shown results support the known role of recombination as an amplifier of the efficacy of both positive and purifying selection (Presgraves, 2005; Betancourt, Welch, and Charlesworth, 2009; Campos et al.,

2014; Castellano et al., 2015). Thus, the recombination rate is the only genomic feature that shows a positive correlation with ω_a and a negative correlation with ω_{na}. Therefore, genes in high recombination regions experience more efficient purifying and positive selection. Previous studies have demonstrated that the reduction in linkage disequilibrium between nucleotide sites induced by recombination allows them to behave independently, increasing the efficiency of selection acting on them. According to the HRi, a reduced efficacy of purifying and positive selection is expected and observed in regions with low rates of recombination and linkage disequilibrium between sites (Castellano et al., 2015). Recombination rate and the estimates of the HRi effect, in particular, should not be overlooked, but be included as parameters to estimate in population genomic studies (Castellano et al., 2015; Casillas and Barbadilla, 2017). It is of particular interest to know which is the recombination threshold above which HRi disappears, as well as to know the percentage of adaptive mutations that are lost in other species.

4.2.2. Main determinant of the evolutionary rate: gene expression

One of the most important determinants of protein evolution is the expression bias, a measure of how much the expression of a gene is restricted to one or few developmental stages (or tissues). The metric τ has been used for estimating the expression bias for each gene. This metric has been demonstrated to be the best overall index to measure expression specificity (Kryuchkova-Mostacci and Robinson-Rechavi, 2017).

Ubiquitously expressed genes have been found to evolve more slowly than genes with a more restricted expression, which exhibited faster evolutionary rates due to the accumulation of adaptive and non-adaptive substitutions. Larracuente et al. (2008) investigated whether this effect is driven by genes that are expressed in male reproductive tissues (which are known to be fast evolving in several species, including *D. melanogaster*), which was not the case. Their greater propensity to adaptation could be partly due to the fact that these genes are involved in less cellular processes than ubiquitously expressed genes, leading to a less extensive pleiotropy (Larracuente et al., 2008).

However, a negative correlation between the expression bias and the expression level is also found: genes that are broadly expressed exhibit

higher expression levels and *vice versa* for restricted-expressed genes. Therefore, the constraint could be explained by either a high expression level or a low expression bias or a combination of both (Subramanian and Kumar, 2004).

Because this strong relationship between the levels of constraint and expression is found in yeast (Pal et al., 2001), which is not affected by the expression bias since yeast is a unicellular organism, it is reasonable to hypothesize that gene expression level is the first evolutionary determinant of the adaptive rate and not the expression bias (Subramanian and Kumar, 2004). Expression bias, created by an increased complexity in terms of the number of tissues and developmental stages of more complex organisms, can be considered a modifying factor of the effect of purifying selection (Subramanian and Kumar, 2004).

Additionally, a negative correlation between the spatial pleiotropy and *Fop* was observed (Figure 3.18C), which indicates that ubiquitously expressed genes have a higher codon usage bias (measured with *Fop*) than narrowly expressed ones. This is in agreement with a selective pressure on codon usage depending not only the level of expression –because highly expressed genes tend to have a higher codon bias (Quax et al., 2015)–, but also on the number of body parts in which genes are expressed as suggested in Duret and Mouchiroud (1999).

Three complementary hypotheses trying to explain why highly expressed genes evolve slowly have been proposed (reviewed in Rocha, 2006): (i) the functional hypothesis states that highly expressed proteins require more cell resources and probably control more important processes –thus, being less prone to change; (ii) the translation accuracy hypothesis states that the efficiency of translation of proteins depends on its tRNA distribution and therefore the rate of synonymous substitutions will be reduced to favor the maintenance optimal codons; (iii) the translational robustness hypothesis states that selection will constraint amino acid changes affecting misfolding or mistranslations.

4.2.3. Intron length orchestrates expression levels

A positive correlation between the intron length and expression bias was found (Figure 4.4), suggesting that housekeeping genes tend to have shorter introns than restricted-expressed genes. Some authors propose

that housekeeping genes require very little regulation and thus, their introns are shorter (the *genomic design hypothesis*, Eisenberg and Levanon, 2003). The contrary happens with stage- or tissue-specific genes, which need more regulatory elements, which can be located within introns (Castillo-Davis et al., 2002). Other authors claim that short introns are the result of selection for translation efficiency because short introns would reduce the cost of transcription in highly expressed genes (the *economy selection hypothesis*, Castillo-Davis et al., 2002; Rao et al., 2010). An alternative to this hypothesis argues that is not the energetic cost, but the transcription time, a more important factor accounting for the small introns of some genes (the *time-cost hypothesis*, Chen et al., 2005).

Genes that are expressed in the early moments of the embryonic development requires an exact timing of gene expression to ensure the proper development of the embryo (Swinburne and Silver, 2008; Artieri and Fraser, 2014). Thus, the time-cost theory might be the most feasible explanation of their short introns. As it is shown in Figure 3.11E, *D. melanogaster* genes expressed in the first two hours of development have the shortest introns of the whole life cycle (also reported in Anderson, 1973; Artieri and Fraser, 2014; Heyn et al., 2014). This short intron size would be imposed by the very fast cell divisions occurring in those hours. Since cells divide very rapidly in early *D. melanogaster* development, there is no time to transcribe, splice and translate long genes and genes with long introns. In vertebrates, the earliest stages of development involve also fast cell divisions and a delayed start of major embryo patterning through cell signaling (O'Farrell, Stumpff, and Su, 2004; Heyn et al., 2014; Siefert, Clowdus, and Sansam, 2015). However, Heyn et al. (2015) pointed out that not necessarily all genes expressed during embryogenesis have short introns.

In the next hours of the *D. melanogaster* embryo development the intron length of genes expressed in these stages appeared to be longer (Figure 3.11E). In 1986, Gubb introduced the *intron delay hypothesis*, stating that intron length could function as a time delayer and aid the orchestration of gene expression patterns. A delayed expression can create oscillating patterns of gene expression. In mice, the *Hes7* gene is an example of this phenomenon. *Hes7* is involved in the somite segmentation during embryo development. Removal of its introns results in an earlier expression (19 minutes), linked to severe developmental defects (Takashima et al., 2011). Another study shows that in six *Drosophila* species the expression of zygotic genes with long introns is delayed compared to shorter zy-

gotic genes (Artieri and Fraser, 2014). These observations indicate that intron length delay plays an important role in regulating gene expression during development.

4.2.4. An intergenic Hill-Robertson interference

As previous studies found, short genes tend to accumulate both adaptive and non-adaptive substitutions (Comeron, Williford, and Kliman, 2008; Larracuente et al., 2008). Thus, natural selection is less effective on short protein-coding genes than on long ones. A possible explanation for this observation could be the fact that for a given adaptive mutation rate, longer genes would have more adaptive segregating sites competing against each other in different haplotypes. This would produce a kind of intergenic or intraexonic Hill-Robertson interference (Hill and Robertson, 1966). This proposed idea is in agreement with the observed negative correlation between gene size and *Fop* (Figure 4.4). Li (1987) predicted that the efficacy of selection on optimal codons decreases with increasing gene size. Within a gene, recombination rarely happens, and thus, there is a high degree of linkage between sites. The efficiency of natural selection in purging deleterious or fixing advantageous variants is lower when there is linkage.

4.2.5. Phylogenetic age as a new proxy for gene conservation

By using phylostratigraphic maps to assign a phylogenetic age to each *D. melanogaster* gene, it was shown that phylogenetically older genes are more conserved than phylogenetically recent ones (Domazet-Loso, Brajković, and Tautz, 2007; Domazet-Lošo and Tautz, 2010). The same results have been found in plants (Guo, 2013).

The reasons by which younger genes show higher ω_a and ω values than older genes are complex. Younger genes exhibit higher expression bias and tend to be expressed in less anatomical structures than older genes (Figure 3.18A). This suggests a scenario in which emerging genes start with very restricted expression (both in anatomical space and developmental time) and thus have a low level of pleiotropy that would facilitate their further evolution. On the other hand, older genes are more likely to be metabolic or housekeeping genes with essential functions

that are unlikely to change in an adaptive way (Hastings, 1996; Duret and Mouchiroud, 2000; Daubin and Ochman, 2004; Zhang and Li, 2004; Albà and Castresana, 2005; Domazet-Loso, Brajković, and Tautz, 2007; Wolf et al., 2009a).

There is a concern with the use of phylostratigraphic maps. These maps assign a phylogenetic age to each protein-coding gene in a species of interest, based on the phylogenetic level at which orthologs for that gene are detected. The accuracy of phylostratigraphic inferences relies on BLAST searches, which show some limitations when sequences are highly divergent (Elhaik, Sabath, and Graur, 2005). However, Domazet-Lošo et al. (2017) have shown that phylogenetic data is not biased by BLAST results.

4.2.6. Novel genomic features for evolutionary rate determination

According to current knowledge, the number of exons and transcripts has never been directly associated ω_a or ω_{na}. However, as one would expect, the number of exons positively correlates with gene size (Figure 4.4). Likewise, the greater the number of exons is, the greater the number of different isoforms, which can be produced by re-shuffling exons via alternative splicing. Therefore, the relationship between the number of exons or transcripts with the selective regimes could be due to the natural or physical correlation between the number of exons and transcripts with gene size.

Additionally, the negative correlation between the number of transcripts and gene expression bias suggests that broadly expressed genes code for more isoforms than stage-specific genes. Since broadly expressed gene isoforms must operate in a more diverse set of cell compartments, biological processes or tissues, this correlation is expected. Haerty and Golding (2009) found that genes undergoing alternative splicing (more than one transcript per gene) have a broader pattern of expression, which is associated with a lower divergence in comparison with genes with a single annotated protein isoform (Haerty and Golding, 2009).

4.3. Population genomics at the multiomic level

The ultimate level is the multiomic or integrative level, in which the genomic patterns explained above are correlated with large -omics datasets. In contrast to the genomic sequence, -omics layers, vary during the lifetime or body parts of an individual, representing intermediate phenotypes between the genomic space and the final phenotype on which natural selection ultimately acts (Civelek and Lusis, 2014; Casillas and Barbadilla, 2017).

4.3.1. Towards a population -omics synthesis

Natural selection acts primarily on the phenotypic properties of organisms and only act secondarily on the genotype to the extent that it determines the heritability of these properties of the phenotype. The genomic revolution has led to the currently paradoxical situation in which more information on selection in the genome is available than on the phenotype of the organism. The action of the selection in the whole phenotype of an organism has never been studied to this date, nor has any study integrated both levels of selection on a genomic scale. In this final part, we deal with this global fitness-phenotype-genotype integration, more specifically, to draw an exhaustive map of the selection acting on the complete development of the species *D. melanogaster*. A total of three layers of -omics information have been integrated (Fig-

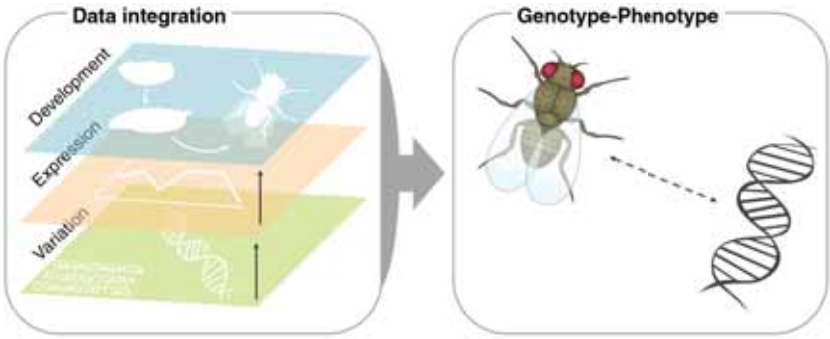


Figure 4.5 Integration of three -omics layers. The first layer is the genomic variation data in *D. melanogaster*. The second is the developmental transcriptome. The third one, gene expression data with accurate annotations of the anatomical regions and timing in which genes are expressed.

ure 4.5). The first layer is the genomic variation data in *D. melanogaster* that has extendedly been analyzed, demonstrating the omnipresence of the selection (Mackay et al., 2012; Huang et al., 2014). Secondly, the advent of NGS technologies has boosted the breadth of available functional datasets. The Berkeley *Drosophila* Genome Project (BDGP) has generated a database of gene expression with accurate annotations of the anatomical regions in which genes are expressed in six embryo developmental stages (Tomancak et al., 2002, 2007). So far, the expression of 8,405 genes have been documented with 137,115 digital photographs (BDGP insitu; `http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl`; last accessed: October 2018). And third, the modENCODE project (Graveley et al., 2011) provides the most complete gene expression database through *D. melanogaster* complete life cycle (it includes 17,788 genes over most developmental and life cycle stages).

Two analyses at different scales were performed. The first analysis consisted of the measurement of the action of natural selection in the genes expressed across the life cycle development of *D. melanogaster*, providing a temporal view. In the second analysis, natural selection was mapped across different anatomical organs during six embryo developmental stages, providing a spatio-temporal view.

4.3.2. Measuring the action of natural selection across the *D. melanogaster* life cycle

Through the integration of population genomics data with the developmental transcriptome of *D. melanogaster* three main conclusions can be drawn. First, the rate of adaptive substitution measured along the life cycle of *D. melanogaster* reveals two peak periods: one encompassing the four initial hours of the embryonic development and one encompassing from the L3 larval stage onwards. Second, the pattern of the selection statistics measured over development mirrors that of the genetic features analyzed. And third, our results support the hourglass model of development evolution.

Adaptive evolution mainly acts in the adult

D. melanogaster is a holometabolous insect with an indirect development, meaning that its development includes four life stages –embryo, larva, pupa and adult– with two active free-roaming phases (larva and adult) and two sessile developmental phases (embryo and pupa; Bainbridge and Bownes, 1981).

The larval and adult phenotypes, especially their morphology, arise primarily through the genetic, cellular and tissue interactions of embryonic and pupal development (metamorphosis), respectively. Therefore, it is hypothesized that adaptation occurring in the larva or in the adult should be reflected not only in the genes expressed in the larva or in the adult but also in those expressed in the embryo (for the larva) and the pupa (for the adult). The observation that genes expressed in mid and late embryonic development show lower rates of non-synonymous substitutions than genes expressed in the pupal stages suggests that adaptation has occurred preferentially in the adult rather than in the larva.

The differences in the adaptation rates found for the sets of genes expressed in male and female adults lead to the following conclusions. The fact that males exhibit higher adaptive rates can be accounted by male specific processes occurring mostly after the determination of the adult morphology during pupation. In fact, male adult genes exhibit a high expression bias (Figure 3.11F) and higher values of ω and ω_a compared to female-biased genes (Figure 4.6). A GO-enrichment reveals an overrepresentation of terms related to *post-mating behaviour* (fold enrichment=7.25, FDR=3.63×10^{-3}), *sperm storage* (fold enrichment=5.93, FDR=8.26×10^{-3}) or *flagellated sperm motility* (fold enrichment=3.57, FDR= 4.76×10^{-3}). In that line it has previously been reported in *D. melanogaster* that the genes exhibiting the highest rates of adaptive change are involved in male reproductive processes, such as spermatogenesis (Civetta and Singh, 1998; Swanson et al., 2001; Artieri and Singh, 2010) and immunity (Schlenke and Begun, 2003; Jiggins and Kim, 2007; Obbard et al., 2009; Early et al., 2017) and are male-biased (Pröschel, Zhang, and Parsch, 2006; Baines et al., 2008). On the other hand, female-biased genes appear to have higher expression than males-biased ones (Figure 3.11H) which can explain the high purifying selection acting on them (Figures 3.9 and 4.6). Because immune and male reproductive genes could account for false positive signals found in our results, they should be excluded of the analyses

to avoid possible bias and confounding effects (Larracuente et al., 2008; Castellano et al., 2015).

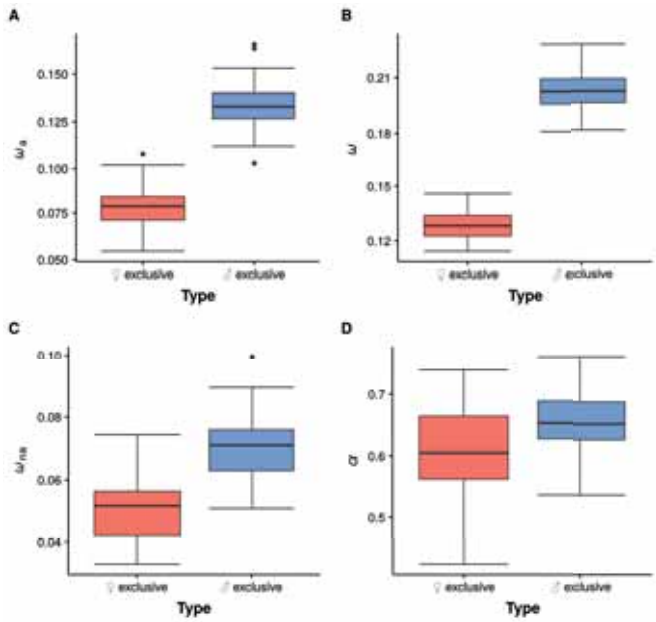


Figure 4.6 Female- and male-biased genes. A gene is considered female-biased if it is expressed during female adulthood but not in any male adult stage, and vice versa for male biased genes. A gene is considered as expressed if RPKM $\geqslant$ 10 (high stringent criteria). **A**. ω_a between female and male genes. **B**. ω between female and male genes. **C**. ω_{na} between female and male genes. **D**. α between female and male genes. Each boxplot (100 bootstrap replicates per category) in a plot is calculated for a randomly drawn sample of the set of genes in each category with replacement. Number of female-biased genes: 507. Number of male-biased genes: 688.

In contrast to adult morphology, which arises primarily through developmental processes occurring in the previous pupal developmental stages, spermatogenesis and immune response are continuous processes that occur mostly through the whole adult lifespan and that are regulated by genes expressed in the adult. In contrast to a previous report (Artieri, Haerty, and Singh, 2009), male adults do not show higher rates of non-adaptive substitutions than the pupa, but rather show similar levels (Figure 3.9). In another previous study (Davis, Brandman, and Petrov, 2005), it was found that genes with the highest number of non-synonymous substitutions are more intensively expressed in the larva and pupa than in the embryo and biased towards male adults. These results are consistent with the findings presented here.

An important novelty of our approach is that by using DFE-alpha, adaptive and non-adaptive evolution can be assessed independently. This allows the inference that the lower level of conservation in early pupal and male stages is due to adaptive non-synonymous substitutions and not to non-adaptive non-synonymous substitutions. In the earliest developmental stages instead, the lower sequence conservation is due mostly to non-adaptive substitutions. Most likely the latter is due to maternal-effect genes. Previous studies have already pointed out that selection in maternal genes is less efficient (although these studies do not relate that to the lower conservation of early development, Cruickshank and Wade, 2008; Fairbanks, 2010). This is the case because in females the alleles in the loci of maternal genes can affect the fitness of the offspring, while it is not the case for males. From a population genomic perspective, selection is only half as strong when acting on maternal-genes that on zygotic-genes (Wade, Priest, and Cruickshank, 2009). Therefore, many non-adaptive variants cannot be eliminated from the population (leading to the higher ω_{na}, Figure 3.9C). This high rate of non-synonymous non-adaptive substitutions in the maternal genes is consistent with recent findings that indicate that there is some variation in which genes are expressed in the earliest developmental stages. Thus, while the maternal genes involved in the earliest embryo patterning can be different within Diptera, the zygotic genes expressed right after (so-called GAP genes), tend to be the same in all the Diptera species analyzed so far (Wotton et al., 2015).

Genomic features mirror the patterns of selective regimes

The pattern of the selective regimes measured over the life cycle mirrors that of the pattern of the genomic features analyzed. Thus, the highly conserved mid and late embryonic development stages express genes that, on average, are larger, have more exons, more isoforms and larger introns.

The correlations between the genomic features and the population statistics have already been discussed (section 4.2), but their distribution over developmental and life cycle stages has not been analyzed before. With this statistical analysis, it is not possible to conclude whether the temporal adaptation pattern is a consequence of differences in genomic features over the life cycle (thus, selective effects on developmental stages become secondary), or if these genomic features are a consequence of

differential adaptation over the life cycle. However, there are some intrinsic characteristics of development that make the first option more likely, which are addressed in the next paragraphs.

In *D. melanogaster* (Salvador-Martínez and Salazar-Ciudad, 2015) and in the ascidian *Ciona intestinalis* (Salvador-Martínez and Salazar-Ciudad, 2017) the area of expression of genes over the embryo's anatomy decreases quantitatively over developmental time. To understand why this is the case, one has to consider that the spatial information in the embryo is built over the life cycle, especially during embryonic and pupal development (Salazar-Ciudad, Jernvall, and Newman, 2003). Spatial information means the information on where each organ, tissue or cell type is located in the embryo's anatomy. This spatial information starts being small (e.g., in the zygote) and progressively increases as specific cells, tissues and organs form in specific parts of the body. The same occurs for the spatial information at the level of gene expression: genes start being expressed in wide areas of the embryo (e.g., in the part that will become the thorax of the fly) and progressively become restricted to smaller areas of it (e.g., in parts that will become the flight muscles of the second thoracic segment). This trend is not only observed during embryonic development but throughout all the life cycle (as Figure 3.11F shows). This spatial restriction of gene expression over time is an intrinsic property of development. It is a consequence of spatial information having to be built from the zygote and has no direct adaptive advantage *per se*.

From this perspective, the temporal pattern of the measured selection parameters would be a consequence of genes becoming more restrictedly expressed, while their level of expression decreases. The earliest development would escape this trend due to the lower efficiency of natural selection on maternal genes. The same argument used for time and expression level can be applied to time and expression bias. Late development and post-embryonic stages have higher expression bias. Genes with higher expression bias tend to be less conserved and are more likely to exhibit adaptive non-synonymous substitutions. Thus, the temporal pattern of the measured selection parameters could be a consequence of early developmental genes being expressed widely in space and time, which promotes conservation, and late development genes being expressed in a more restricted manner in space and time, that facilitates adaptive non-synonymous substitutions.

This argument does not directly explain why the temporal pattern of selection statistics correlates with the number of exons, number of transcripts, intron length and gene size. An alternative and largely complementary explanation would be that, as suggested from a more qualitative evo-devo perspective (Kennison, 1993; Gellon and McGinnis, 1998), embryonically expressed genes have a more complex regulation than post-embryonically expressed genes –this is also discussed in section 4.2.3. Although the former genes are expressed in wider areas of the embryo, their expression changes more in time and space than that of post-embryonically expressed genes. This more complex regulation may require a more complex genetic structure, such as manifested by larger genes having more exons, more transcripts and larger introns. The larger intron length of developmental genes may also be a reflection of complex regulation but at the level of *cis*-regulatory elements, since *cis*-regulatory elements can be located within introns, too (see section 4.2.3). The larger area of expression, less temporally restricted expression and more complex gene structure may also reflect that mid and late developmental genes interact with more other genes than genes expressed later. This would make them, in rough terms, more pleiotropic and thus, less likely to change.

The hourglass model of development

The pattern of adaptation and constraint through the development is roughly consistent with the hourglass model but not with the von Baer's law. However, the fit to the hourglass model is rather weak, since there are no major differences in ω between embryonic stages after the eighth hour (Figure 3.9), except for genes in cluster 8 (Figure 3.10). Cluster 8 contains genes specifically expressed in the last hours of the development. During the first 2 hours, ω is significantly high (permutation test, hours 0–2: p-value = 0.032), but from hours 6–8 to hours 22–24 ω is lower than expected based on the permutation test (p-value < 0.001). This is also the case for ω_a. In contrast with previous studies (Kalinka et al., 2010; Levin et al., 2016), this study does not show that the latest stages of embryonic development are less conserved. However, genes whose expression is restricted only in late embryonic development (cluster 8, Figure 3.10), show a significant high ω and marginally significant high ω_a. These genes are only a small proportion of the genes expressed in the last embryonic developmental stages and thus, have a minor effect on our calculations of ω_a, ω and α in these stages (thus likely explaining

the differences between our study and Kalinka et al. 2010). The difference in d_N between mid- and late-developmental stages in Kalinka et al. (2010) is however rather subtle, too. Overall, results are compatible with Kalinka et al. (2010).

The hourglass model was proposed on the basis of knowledge about *D. melanogaster* and vertebrate's development (Slack, Holland, and Graham, 1993; Duboule, 1994; Raff, 1996). The life cycles of the fly and the mouse are quite different. Mice, as all amniotes, are direct developers, meaning that development gives rise to a juvenile and later, gradually, to an adult. Fruit flies are indirect developers in which embryonic development gives rise to a free-roaming larva and, by a rather abrupt process of metamorphosis, gives to an adult. If the hourglass model is understood for the whole life cycle the results are roughly consistent with it at the genetic level: genes expressed during embryonic development are highly conserved, except for the genes expressed in the earliest stages, while the genes expressed later, from the larval stage L3 onwards, show less conservation and more adaptation. On the other hand, this temporal hourglass pattern can also be understood as development generally obeying von Baer's law, but departing from it in the earliest stages. It is hypothesized that this departure would arise from the lower efficiency of selection on maternal genes and as a consequence of the reduced gene structure complexity required for fast nuclei divisions in early development.

4.3.3. Mapping natural selection through the embryo's anatomy

This work measures which parts of the embryo's body exhibit significantly higher (or lower) adaptation levels (measured with ω_a) or constraint (measured with ω_{na}), compared with the rest of genes expressed in the other anatomical structures of the embryo. The anatomical structures with high ω_a values should be interpreted as body regions with high rates of adaptive substitutions. The ones with high ω or ω_{na} should be interpreted as body regions under relaxed natural selection, whereas the anatomical structures with low ω or ω_{na} values should be interpreted as body regions under a history of selective constraint.

The latest embryonic stage analyzed, stage 13–16, shows the highest number of anatomical terms exhibiting evidence of selection, both adap-

tation and selective constraint. In this stage, the anatomical structures and the gene spatial co-expression patterns are positioned and shaped in very similar ways to those of the larva (no major morphogenetic rearrangements occur from that stage onwards, Hartenstein, 1993). In that sense, the results in this latest stage could be taken as a proxy for adaptation over the body parts of the functional larva.

In summary, high rates of adaptive substitution are found the "Germ line", the "Garland cells/Plasmatocytes/Ring gland", and also the "Head mesoderm/Circulatory system/Fat body." Most of the rest of the body seems to be under selective constraint. Results are consistent with previous findings from other non-development studies. Thus, the evidences of adaptation in the "Germ line" is consistent with previously reported high K_a and α in testis- or sperm-specific genes (Civetta and Singh, 1995; Wyckoff, Wang, and Wu, 2000; Meiklejohn et al., 2003; Pröschel, Zhang, and Parsch, 2006; Haerty et al., 2007; Assis, Zhou, and Bachtrog, 2012) and sperm-related genes already expressed in germ line cells (Civetta et al., 2006; Bauer DuMont et al., 2007).

The category "Garland cells/Plasmatocytes/Ring gland" is closely linked to the immune system. Plasmatocytes comprise the 95% of all the immune cells in *D. melanogaster* (similar to human macrophages, Ratheesh, Belyaeva, and Siekhaus, 2015). The ring gland has also been related to the immune system (Christesen et al., 2017). As explained above, the immune system has already been shown to exhibit high rates of adaptive substitutions (high α) in *D. melanogaster*.

Overall, the results suggest that there is a high degree of conservation in genes expressed over most parts of the embryonic anatomy. In particular, ectodermal-derived organs. This is in agreement with the observation that the ectoderm is the oldest germ layer in *D. melanogaster* (Domazet-Loso, Brajković, and Tautz, 2007). Adaptive substitutions are found in the set of genes expressed in anatomical structures involved in reproduction and immunity.

Also, evidence of relaxation of selection was found in the first stage, supporting the results found with the modENCODE data in the temporal analysis. In overall, the results of the temporal analysis and the spatio-temporal analysis are consistent and complementary.

Taken all together, it can be stated that selective constraint is pervasive over most of the embryo's anatomy, except for anatomical structures

that show evidence of adaptation in the adult (immune system and reproductive-related genes) and a relaxation in the first stage due to the maternal-effect genes.

Results are neither driven by recombination context nor expression level

The map of selection regimes through the embryo's anatomy does not directly explain why these specific anatomical structures exhibit high rates of adaptive substitutions. However, some additional analyses were performed to limit some possible explanations. For example, it could be that the genes expressed in organs under positive selected are preferably found in regions of high recombination or low genic density. It can also be the case that these genes are low pleiotropic or phylogenetically younger.

First, it was checked if there was a correlation between the expression bias and the anatomical organs exhibiting high adaptation. No relation between these anatomical structures and the level of temporal pleiotropy (expression bias) was found (Figure 3.16C). It was expected that genes expressed in anatomical structures exhibiting the higher level of adaptation have a high expression bias. Furthermore, a measure of the spatial pleiotropy was also incorporated, an index measuring the pleiotropic effects of a gene on the embryonic anatomy, without finding a relation (Figure 3.16D).

The analysis of *Fop* and phylogenetic age revealed that these two features do not explain the adaptation levels found. It is expected that anatomical organs enriched in phylogenetic recent genes and/or with a low *Fop*, are the ones exhibiting higher rates of adaptation, which is not the case.

Finally, the analyses on recombination, mutation and gene density also showed no differences between body parts, i.e, genes expressed in organs under selection are not located in a high recombination context or low genic density regions.

In that respect, these results do not accommodate straight-forward genomic explanations but rather suggest that there may be some functional features of the "Garland cells/Plasmatocytes/Ring gland" and "Germ line" which have favored the accumulation of adaptive substitutions in the genes they express, at least when compared with other parts of the anatomy, regardless of their pleiotropic effects.

Permutation tests as a powerful statistical strategy to detect selection in multiple anatomical parts

Each anatomical structure expresses a particular set of genes. However, some of these genes can be expressed in more than one anatomical structure. The number of genes shared depends on the function of the anatomical structures: those structures having a similar function share a higher number of genes (Figure 4.7). This non-independence between anatomical structures make it difficult to analyze them separately, as they are not completely independent regarding the genes they express. This dependence has further implications for statistical tests and how in these tests related genes are taken into account (see section 2.3.2).

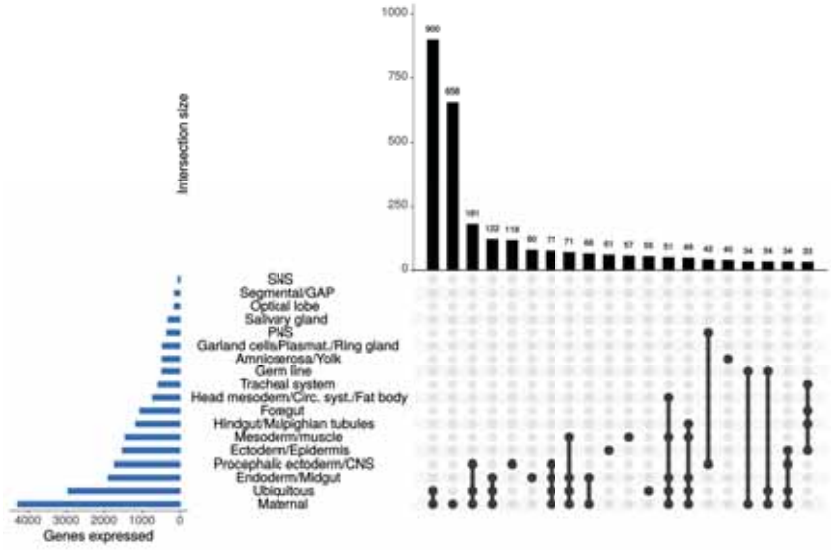


Figure 4.7 Genes shared between anatomical organs. The matrix layout shows intersections of the 18 anatomical organs, from more frequent to less frequent intersections (only the 20 first intersections are shown). The blue bars on the left side represent the number of genes expressed in each anatomical structure, while the black bars on the top represents the number of genes shared between the anatomical organs. Image performed with the R package `UpSetR` (Conway et al., 2017).

Permutation tests are better suited for avoiding statistical Type I errors and are considered a robust alternative to the Bonferroni correction (Sham and Purcell, 2014). One of the main advantages of this method, which has been applied for the first time in this kind of data, is that it can be applied to any statistic and can incorporate distributional and dependence characteristics inherent to the data used, making it a robust

test (Westfall and Young, 1993). Most importantly, when using permutation tests the null distribution is empirical, i.e., is obtained by calculating all possible, or a very large number of, values of the statistic under rearrangements of the labels of the observed data points (Berry, Mielke, and Johnston, 2016). Therefore, in the case of the analyses performed here, the null distribution of adaptive and constraint rates is different for the different analyses, as each one is comprised of a different number and combination of genes.

The case of the salivary glands and GAP genes

There are two anatomical structures that appear in the analyses as an exception to many of the identified trends. The salivary glands express old genes that, in contrast to the rest of the organs derived from the ectoderm, exhibit rather high *Fop*. The high *Fop* may be explainable by both the old age (i.e., more time to optimize their sequence) of the genes expressed and the fact that many of the genes expressed in the salivary glands are well known to be expressed at very high levels (Andrew, Henderson, and Seshaiah, 2000). Genes that are expressed at high levels are known to usually have rather high *Fop*, because this facilitates a faster and efficient translation as explained in section 4.2.2 (Gingold and Pilpel, 2011; Quax et al., 2015).

The "Segmental/GAP" anatomical structure is also exceptional because it expresses the highest proportion of new genes. "Segmental/GAP" genes are expressed very early on before the germ layers form. In fact, previous studies have shown that there is substantial variation between Diptera, in which genes act early in development as segmental and GAP genes (Wotton et al., 2015). Since these genes are all transcriptional factors, it is not surprising that they are all relatively young (old genes tend to be metabolic genes involved in processes that are shared among distantly related groups, Wolf et al., 2009b).

Main caveats about the assumptions of our approach

There are a number of caveats to be considered when using gene expression as a phenotype. First, the amount of adaptive amino acid substitutions in the set of genes expressed in an anatomical structure may not accurately reflect the amount of adaptive phenotypic changes

in it. Development is a complex process with myriads of genetic and cell biomechanic interactions, which lead to a complex relationship between genetic variation and phenotypic variation. It can be, for example, that some anatomical structures show only a small number of adaptive changes in the genes they express and then being not detectable by these methods, but that those genetic changes have comparatively large effects on the phenotype.

Second, only changes in coding regions are considered, although there is plenty of evidence of adaptation resulting from changes in regulatory regions (Davidson, 2001; Carroll, 2005). Some studies conclude that non-coding elements tend to experience more adaptive events than protein-coding genes, at least in mice (Eyre-Walker and Keightley, 2007; Halligan et al., 2013) and *Drosophila* (Kousathanas et al., 2011; Mackay et al., 2012).

Third, variation in a gene can have an effect on anatomical structures where such genes are not expressed (Gilbert, 2014). This is the case of extracellularly diffusible proteins. During development, a signal from a group of cells can influence another adjacent group and this interaction is necessary for cell development and differentiation. Signals can be transmitted in different ways and one of them is through the extracellular space in the form a secreted diffusible protein. Another example is the genes involved in the production of mechanical forces (reviewed in Zhou et al., 2009 and Vining and Mooney, 2017), which are necessary for the correct patterning, growth and morphogenesis of a developing embryo.

Although all these caveats should be kept in mind there is no reason to expect that, a priori, the complexity of the genotype-phenotype map (or for that matter the amount of *cis*- regulation, signaling, or mechanical forces) to be dramatically different between anatomical structures.

4.4. Concluding remarks

The current -omics era is calling for more integrative, multi-level approaches to study adaptation. A novel approach for mapping the phenotypic adaptation and natural selection over the complete anatomy of the embryo of *D. melanogaster* has been presented. The emergence of techniques such as *spatial transcriptomics* offering more resolutive maps

(Stahl et al., 2016, even for the single cell level, Karaiskos et al., 2017; Shah et al., 2018) promises that natural selection will be charted with an unprecedented resolution also outside the *D. melanogaster* species.

Population genomics is concerned with genome variation, but natural selection acts upon the phenotype, not directly on the genotype, and the genomic dimension, albeit necessary, is not sufficient to account for a complete picture of organismal adaptation (Lewontin, 2000; Casillas and Barbadilla, 2017). Population genomics is no longer a theoretical science, it has become an interdisciplinary field where bioinformatics, large functional multiomics datasets, statistical and evolutionary models and emerging molecular techniques are all integrated to get a systemic view of the causes and consequences of evolution. This thesis is a first step towards the final goal of charting a complete fitness-phenotype-genotype map. At this coming moment, population genetics theory will become integrated in a systemic evolutionary theory (Casillas and Barbadilla, 2017).

Chapter 5

CONCLUSIONS

5 Conclusions

The conclusions of this work are the following:

1. The conducted comparison of McDonald and Kreitman test (MKT) methodologies using both empirical and simulated data shows that the method proposed by Mackay et al. (2012), the extended MKT (eMKT), is the one that performs best if single-gene data is used. The eMKT allows to remove the effect of negative selection as well as to quantify it. On the contrary, the asymptotic method developed by Messer and Petrov (2013) is preferred when data is abundant as in the case of concatenated genes.

2. The original MKT is a powerful method to detect recurrent positive selection on coding sequences at the molecular level, granted that slightly deleterious polymorphism is absent. This supposition is unrealistic in *D. melanogaster* and probably in many other species, and therefore other alternative MKT should be applied.

3. The four categories of genomic features analyzed along the *D. melanogaster* genome –i.e., gene architectonic, gene expression, genomic context and gene phylogenetic features– are strongly correlated with both the adaptive and non-adaptive rates of protein-coding genes.

4. Our results support the known role of recombination in reducing the Hill-Robertson interference. The constraint due to purifying selection is positively correlated with gene architectonic complexity, evolutionary age and/or expression levels. The adaptation rate due to positive selection seems restricted to those coding sequences with low structural complexity, evolutionary younger and expressed in a few developmental stages at a low level.

5. The integration of population genomics data with the developmental transcriptome of *D. melanogaster* has allowed us to measure the rate of adaptation and selective constraint through the complete life cycle of the fruit fly.

6. Considered over the whole life cycle, *D. melanogaster* seems to fit the hourglass model of evolutionary development at the molecular level. Genes expressed during mid- and late-embryonic development are highly conserved, while genes expressed in the earliest stages and from the larval stage L3 onwards are highly divergent.

7. The higher sequence divergence observed in the firsts hours of the embryo development is mostly due to the accumulation of non-adaptive substitutions. We hypothesize that this departure would arise from the lower efficiency of selection on maternal-effect genes. Additionally, genes expressed in these first stages have on average the shortest introns, probably due to the need for a rapid and efficient expression during the short cell cycles.

8. The pattern of the selective regimes measured over the life cycle mirrors that of the pattern of the genomic features analyzed. Thus, genes expressed in mid- and late-embryonic developmental stages show the highest sequence conservation and the most complex gene structure: they are larger, consist of more exons and longer introns, encode a large number of isoforms and, on average, are highly expressed.

9. The charted fitness-phenotype-genotype map of adaptation and constraint over the complete anatomy of the embryo of *D. melanogaster* suggests that selective constraint is pervasive over most of the embryo's anatomy, particularly on the digestive and nervous systems, except for the anatomical structures that also show evidence of adaptation in the adult, the immune and reproductive systems, and a relaxation of selection in the first stage due to the maternal-effect genes.

10. A novel permutation test has been applied to infer the departures of adaptation and selective constraint simultaneously for all body parts of the embryo. This permutation test captures the correlational structure of the data and has a higher statistical power compared to standard permutation test procedures.

11. The genes annotated with the anatomical terms "Salivary glands" and "Segmental/GAP" depart from the trends identified. The salivary glands express old genes, that, in contrast to the rest of the ectoderm, exhibit a rather high frequency of optimum codons (*Fop*). The high *Fop* could be explained by both the old age of these genes and the fact that many of them are well known to be expressed at very high levels. In contrast, the genes annotated with the anatomical term "Segmental/GAP" are the youngest compared to the genes annotated in the other anatomical terms.

12. The last embryonic stage analyzed (13–16) exhibits the most contrasting values of both adaptation (ω_a) and constraint (ω) between anatomical structures. In this stage, the anatomical structures and the gene spatial co-expression patterns are positioned and shaped in very similar ways to those of the larva. In that sense, the results in the latest stage could be taken as a proxy for adaptation over the body parts of the functional larva.

13. Genes with a low spatial pleiotropy, i.e., expressed in a few anatomical terms, are evolutionary younger and exhibit higher rates of evolution. In contrast, genes that are highly pleiotropic are phylogenetically older and more evolutionary constrained.

14. By measuring natural selection in genes expressed across development or different body parts, a systemic view of the causes and consequences of evolutionary and functional effects of genomic variation. This thesis is a first step in the pursuit to achieve an ultimately unified fitness-phenotype-genotype map.

Bibliography

Adams, M. D. et al. (2000). "The genome sequence of *Drosophila melanogaster*." *Science* 287.5461, pp. 2185–2195.

Akashi, H. (1999). "Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination." *Genetics* 151.1, pp. 221–238.

Akashi, H. (2003). "Translational selection and yeast proteome evolution." *Genetics* 164.4, pp. 1291–1303.

Albà, M. M. and J. Castresana (2005). "Inverse relationship between evolutionary rate and age of mammalian genes." *Molecular Biology and Evolution* 22.3, pp. 598–606.

Alberch, P. (1980). "Ontogenesis and morphological diversification." *American Zoologist* 20.4, pp. 653–667.

Anderson, D. T. (1973). *Embryology and phylogeny in annelids and arthropods*. Pergamon Press, New York, p. 495. ISBN: 0080170692.

Andolfatto, P. (2001). "Contrasting patterns of X-Linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*." *Molecular Biology and Evolution* 18.3, pp. 279–290.

Andrew, D. J., K. D. Henderson, and P. Seshaiah (2000). "Salivary gland development in *Drosophila melanogaster*." *Mechanisms of Development* 92.1, pp. 5–17.

Arbeitman, M. N., E. E. M. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White (2002). "Gene expression during the life cycle of *Drosophila melanogaster*." *Science* 297.5590, pp. 2270–2275.

Arthur, W. (1977). *The origin of animal body plans: a study in evolutionary developmental biology*. Cambridge University Press, p. 360.

Artieri, C. G. and H. B. Fraser (2014). "Transcript length mediates developmental timing of gene expression across *Drosophila*." *Molecular Biology and Evolution* 31.11, pp. 2879–2889.

Artieri, C. G., W. Haerty, and R. S. Singh (2009). "Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of *Drosophila*." *BMC Biology* 7.1, p. 42.

Artieri, C. G. and R. S. Singh (2010). "Molecular evidence for increased regulatory conservation during metamorphosis, and against deleterious cascading effects of hybrid breakdown in *Drosophila*." *BMC Biology* 8.1, p. 26.

Assis, R., Q. Zhou, and D. Bachtrog (2012). "Sex-biased transcriptome evolution in *Drosophila*." *Genome Biology and Evolution* 4.11, pp. 1189–1200.

Auton, A. et al. (2015). "A global reference for human genetic variation." *Nature* 526.7571, pp. 68–74.

Baer, K. E. von (1828). *Über Entwicklungsgeschichte der Thiere. Beobachtung und Reflexion*. Königsberg, Bei den Gebrüdern Borntträger.

Bainbridge, S. P. and M. Bownes (1981). "Staging the metamorphosis of *Drosophila melanogaster*." *Journal of Embryology and Experimental Morphology* 66, pp. 57–80.

Baines, J. F., S. A. Sawyer, D. L. Hartl, and J. Parsch (2008). "Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in *Drosophila*." *Molecular Biology and Evolution* 25.8, pp. 1639–1650.

Ballard, W. W. (1981). "Morphogenetic movements and fate maps of vertebrates." *American Zoologist* 21.2, pp. 391–399.

Balloux, F. and L. Lehmann (2012). "Substitution rates at neutral genes depend on population size under fluctuating demography and overlapping generations." *Evolution* 66.2, pp. 605–611.

Bastian, F., G. Parmentier, J. Roux, S. Moretti, V. Laudet, and M. Robinson-Rechavi (2008). "Bgee: integrating and comparing heterogeneous transcriptome data among species." In: *Data integration in the life sciences*. Springer, pp. 124–131. ISBN: 9783540698272.

Bastock, R. and D. St Johnston (2008). "*Drosophila* oogenesis." *Current Biology* 18.23, R1082–R1087.

Baudry, E., B. Viginier, and M. Veuille (2004). "Non-African populations of *Drosophila melanogaster* have a unique origin." *Molecular Biology and Evolution* 21.8, pp. 1482–1491.

Bauer DuMont, V. L., H. A. Flores, M. H. Wright, and C. F. Aquadro (2007). "Recurrent positive selection at *Bgcn*, a key determinant of germ line differentiation, does not appear to be driven by simple coevolution with its partner protein Bam." *Molecular Biology and Evolution* 24.1, pp. 182–191.

Bazin, E., S. Glémin, and N. Galtier (2006). "Population size does not influence mitochondrial genetic diversity in animals." *Science* 312.5773, pp. 570–572.

Begun, D. J. and C. F. Aquadro (1993). "African and North American populations of *Drosophila melanogaster* are very different at the DNA level." *Nature* 365.6446, pp. 548–550.

Begun, D. J. et al. (2007). "Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*." *PLoS Biology* 5.11, e310.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300.

Berry, A. J., J. W. Ajioka, and M. Kreitman (1991). "Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection." *Genetics* 129.4, pp. 1111–1117.

Berry, K. J., P. W. Mielke, and J. E. Johnston (2016). *Permutation statistical methods*. Springer. ISBN: 9783319287683.

Betancourt, A. J. and D. C. Presgraves (2002). "Linkage limits the power of natural selection in *Drosophila*." *Proceedings of the National Academy of Sciences of the United States of America* 99.21, pp. 13616–13620.

Betancourt, A. J., J. J. Welch, and B. Charlesworth (2009). "Reduced effectiveness of selection caused by a lack of recombination." *Current Biology* 19.8, pp. 655–660.

Bierne, N. and A. Eyre-Walker (2004). "The genomic rate of adaptive amino acid substitution in *Drosophila*." *Molecular Biology and Evolution* 21.7, pp. 1350–1360.

Booker, T. R., B. C. Jackson, and P. D. Keightley (2017). "Detecting positive selection in the genome." *BMC Biology* 15.1, p. 98.

Boyko, A. R. et al. (2008). "Assessing the evolutionary impact of amino acid mutations in the human genome." *PLoS Genetics* 4.5, e1000083.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl (2002). "The cost of inbreeding in *Arabidopsis*." *Nature* 416.6880, pp. 531–534.

Bustamante, C. D. et al. (2005). "Natural selection on protein-coding genes in the human genome." *Nature* 437.7062, pp. 1153–1157.

Campos, J. L., D. L. Halligan, P. R. Haddrill, and B. Charlesworth (2014). "The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*." *Molecular Biology and Evolution* 31.4, pp. 1010–1028.

Canty, A. and B. D. Ripley (2017). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-20.

Caracristi, G. and C. Schlötterer (2003). "Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles." *Molecular Biology and Evolution* 20.5, pp. 792–799.

Carneiro, M., F. W. Albert, J. Melo-Ferreira, N. Galtier, P. Gayral, J. A. Blanco-Aguiar, R. Villafuerte, M. W. Nachman, and N. Ferrand (2012). "Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome." *Molecular Biology and Evolution* 29.7, pp. 1837–1849.

Carroll, S. B. (2005). "Evolution at two levels: on genes and form." *PLoS Biology* 3.7, e245.

Casillas, S., A. Barbadilla, and C. M. Bergman (2007). "Purifying selection maintains highly conserved noncoding sequences in *Drosophila*." *Molecular Biology and Evolution* 24.10, pp. 2222–2234.

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

Casillas, S. and A. Barbadilla (2017). "Molecular population genetics." *Genetics* 205.3, pp. 1003–1035.

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

Castellano, D., J. James, and A. Eyre-Walker (2017). "Nearly neutral evolution across the *Drosophila melanogaster* genome." *bioRxiv*, p. 212779.

Castellano, D., M. Coronado-Zamora, J. L. Campos, A. Barbadilla, and A. Eyre-Walker (2015). "Adaptive evolution is substantially impeded by Hill-Robertson interference in *Drosophila*." *Molecular Biology and Evolution* 33.2, pp. 442–455.

Castillo-Davis, C. I. and D. L. Hartl (2002). "Genome evolution and developmental constraint in *Caenorhabditis elegans*." *Molecular Biology and Evolution* 19.5, pp. 728–735.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov (2002). "Selection for short introns in highly expressed genes." *Nature Genetics* 31.4, pp. 415–418.

C. elegans Sequencing Consortium (1998). "Genome sequence of the nematode *C. elegans*: a platform for investigating biology." *Science* 282.5396, pp. 2012–2018.

Celniker, S. E. et al. (2009). "Unlocking the secrets of the genome." *Nature* 459.7249, pp. 927–930.

Charlesworth, B. (1994). "The effect of background selection against deleterious mutations on weakly selected, linked variants." *Genetical Research* 63.3, pp. 213–227.

Charlesworth, B. and D. Charlesworth (2017). "Population genetics from 1966 to 2016." *Heredity* 118.1, pp. 2–9.

Charlesworth, B., M. T. Morgan, and D. Charlesworth (1993). "The effect of deleterious mutations on neutral molecular variation." *Genetics* 134.4, pp. 1289–1303.

Charlesworth, B. (2010). "Molecular population genomics: a short history." *Genetics Research* 92.5-6, pp. 397–411.

Charlesworth, J. and A. Eyre-Walker (2008). "The McDonald-Kreitman test and slightly deleterious mutations." *Molecular Biology and Evolution* 25.6, pp. 1007–1015.

Chen, J., M. Sun, J. D. Rowley, and L. D. Hurst (2005). "The small introns of antisense genes are better explained by selection for rapid transcription than by "genomic design"." *Genetics* 171.4, pp. 2151–2155.

Christesen, D., Y. T. Yang, J. Somers, C. Robin, T. Sztal, P. P. Batterham, and T. Perry (2017). "Transcriptome analysis of *Drosophila melanogaster* third instar larval ring glands points to novel functions and uncovers a cytochrome p450 required for development." *G3* 7.2, pp. 467–479.

Civelek, M. and A. J. Lusis (2014). "Systems genetics approaches to understand complex traits." *Nature Reviews Genetics* 15.1, pp. 34–48.

Civetta, A. and R. S. Singh (1995). "High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species." *Journal of Molecular Evolution* 41.6, pp. 1085–1095.

Civetta, A. and R. S. Singh (1998). "Sex-related genes, directional sexual selection, and speciation." *Molecular Biology and Evolution* 15.7, pp. 901–909.

Civetta, A., S. A. Rajakumar, B. Brouwers, and J. P. Bacik (2006). "Rapid evolution and gene-specific patterns of selection for three genes of spermatogenesis in *Drosophila*." *Molecular Biology and Evolution* 23.3, pp. 655–662.

Clark, A. G. et al. (2007). "Evolution of genes and genomes on the *Drosophila* phylogeny." *Nature* 450.7167, pp. 203–218.

Clark, K., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers (2016). "GenBank." *Nucleic Acids Research* 44.D1, pp. D67–D72.

Comeron, J. M., M. Kreitman, and M. Aguadé (1999). "Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*." *Genetics* 151.1, pp. 239–249.

Comeron, J. M., A. Williford, and R. M. Kliman (2008). "The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations." *Heredity* 100.1, pp. 19–31.

Comeron, J. M. and M. Kreitman (2000). "The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces." *Genetics* 156.3, pp. 1175–1190.

Comeron, J. M., R. Ratnappan, and S. Bailin (2012). "The many landscapes of recombination in *Drosophila melanogaster*." *PLoS Genetics* 8.10, e1002905.

Connallon, T. and A. G. Clark (2014). "Balancing selection in species with separate sexes: insights from Fisher's geometric model." *Genetics* 197.3, pp. 991–1006.

Consortium, modENCODE (2010). "Identification of functional elements and regulatory circuits by *Drosophila* modENCODE." *Science* 330.6012, pp. 1787–1797.

Consortium, modENCODE et al. (2010). "Identification of functional elements and regulatory circuits by *Drosophila* modENCODE." *Science* 330.6012, pp. 1787–97.

Conway, J. R., A. Lex, N. Gehlenborg, and J. Hancock (2017). "UpSetR: an R package for the visualization of intersecting sets and their properties." *Bioinformatics* 33.18, pp. 2938–2940.

Corbett-Detig, R. B. and D. L. Hartl (2012). "Population genomics of inversion polymorphisms in *Drosophila melanogaster*." *PLoS Genetics* 8.12. Ed. by H. S. Malik, e1003056.

Corbett-Detig, R. B., D. L. Hartl, and T. B. Sackton (2015). "Natural selection constrains neutral diversity across a wide range of species." *PLoS Biology* 13.4. Ed. by N. H. Barton, e1002112.

Coronado-Zamora, M., J. Murga-Moreno, S. Hervás, S. Casillas, and A. Barbadilla (in prep.). "Comparison of five McDonald and Kreitman test approaches using *Drosophila melanogaster* and human population data."

Coronado-Zamora, M., I. Salvador-Martínez, D. Castellano, A. Barbadilla, and I. Salazar-Ciudad (submitted). "Adaptation and selective constraint throughout *Drosophila melanogaster* life cycle."

Cruickshank, T. and M. J. Wade (2008). "Microevolutionary support for a developmental hourglass: gene expression patterns shape sequence

variation and divergence in *Drosophila*." *Evolution and Development* 10.5, pp. 583–590.

Cutter, A. D. (2008). "Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate." *Molecular Biology and Evolution* 25.4, pp. 778–786.

Darwin, C. (1872). *The origins of species by means of natural selection*. John Wanamaker.

Daubin, V. and H. Ochman (2004). "Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*." *Genome Research* 14.6, pp. 1036–1042.

David, J. R. and P. Capy (1988). "Genetic variation of *Drosophila melanogaster* natural populations." *Trends in Genetics* 4.4, pp. 106–111.

Davidson, E. H. (2001). *Genomic regulatory systems*. Elsevier, pp. 63–102. ISBN: 9780122053511.

Davis, J. C., O. Brandman, and D. A. Petrov (2005). "Protein evolution in the context of *Drosophila* development." *Journal of Molecular Evolution* 60.6, pp. 774–785.

Dezso, Z. et al. (2008). "A comprehensive functional analysis of tissue specificity of human gene expression." *BMC Biology* 6.1, p. 49.

Dobzhansky, T. (1955). "A review of some fundamental concepts and problems of population genetics." *Cold Spring Harbor Symposia on Quantitative Biology* 20.0, pp. 1–15.

Dobzhansky, T. and A. H. Sturtevant (1938). "Inversions in the chromosomes of *Drosophila pseudoobscura*." *Genetics* 23.1.

Dobzhansky, T. (1970). *Genetics of the evolutionary process*. Columbia University Press, p. 505. ISBN: 0231083068.

Domazet-Loso, T., J. Brajković, and D. Tautz (2007). "A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages." *Trends in Genetics* 23.11, pp. 533–539.

Domazet-Lošo, T. and D. Tautz (2010). "A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns." *Nature* 468.7325, pp. 815–818.

Domazet-Lošo, T., A.-R. Carvunis, M. M. Albà, M. S. Šestak, R. Bakaric, R. Neme, and D. Tautz (2017). "No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution." *Molecular Biology and Evolution* 34.4, pp. 843–856.

Drost, H.-G. (2014). "A framework to perform phylotranscriptomics analyses for evolutionary developmental biology research."

Drost, H.-G., A. Gabel, I. Grosse, and M. Quint (2015). "Evidence for active maintenance of phylotranscriptomic hourglass patterns in ani-

mal and plant embryogenesis." *Molecular Biology and Evolution* 32.5, pp. 1221–1231.

Drost, H.-G., P. Janitza, I. Grosse, and M. Quint (2017). "Cross-kingdom comparison of the developmental hourglass." *Current Opinion in Genetics & Development* 45, pp. 69–75.

Drummond, D. A., A. Raval, and C. O. Wilke (2006). "A single determinant dominates the rate of yeast protein evolution." *Molecular Biology and Evolution* 23.2, pp. 327–337.

Drummond, D. A. and C. O. Wilke (2008). "Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution." *Cell* 134.2, pp. 341–352.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold (2005). "Why highly expressed proteins evolve slowly." *Proceedings of the National Academy of Sciences of the United States of America* 102.40, pp. 14338–14343.

Duboule, D. (1994). "Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony." *Development* 1994.Supplement, pp. 135–142.

Duchen, P., D. Zivkovic, S. Hutter, W. Stephan, and S. Laurent (2013). "Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population." *Genetics* 193.1, pp. 291–301.

Duret, L. and D. Mouchiroud (1999). "Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*." *Proceedings of the National Academy of Sciences of the United States of America* 96.8, pp. 4482–4487.

Duret, L. and D. Mouchiroud (2000). "Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate." *Molecular Biology and Evolution* 17.1, pp. 68–70.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis." *Bioinformatics* 21.16, pp. 3439–3440.

Early, A. M., J. R. Arguello, M. Cardoso-Moreira, S. Gottipati, J. K. Grenier, and A. G. Clark (2017). "Survey of global genetic diversity within the *Drosophila* immune system." *Genetics* 205.1.

Egea, R., S. Casillas, and A. Barbadilla (2008). "Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing

different classes of DNA sites." *Nucleic Acids Research* 36.Web Server, W157–W162.

Eisenberg, E. and E. Y. Levanon (2003). "Human housekeeping genes are compact." *Trends in Genetics* 19.7, pp. 362–365.

Elhaik, E., N. Sabath, and D. Graur (2005). "The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence." *Molecular Biology and Evolution* 23.1, pp. 1–3.

Elinson, R. P. (1987). "Change in developmental patterns: embryos of amphibians with large eggs." In: *Development as an evolutionary process*. Alan R. Liss, New York, pp. 1–21.

Ermakova, E. O., R. N. Nurtdinov, and M. S. Gelfand (2006). "Fast rate of evolution in alternatively spliced coding regions of mammalian genes." *BMC Genomics* 7, p. 84.

Erwin, D., J. Valentine, and D. Jablonski (1997). "The origin of animal body plans." *American Scientist* 85.2, pp. 126–137.

Eyre-Walker, A. (2006). "The genomic rate of adaptive evolution." *Trends in Ecology & Evolution* 21.10, pp. 569–575.

Eyre-Walker, A. and P. D. Keightley (2007). "The distribution of fitness effects of new mutations." *Nature Reviews Genetics* 8.8, pp. 610–618.

Eyre-Walker, A. and P. D. Keightley (2009). "Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change." *Molecular Biology and Evolution* 26.9, pp. 2097–108.

Fairbanks, L. A. (2010). "Maternal effects in ontogeny and evolution." *Journal of Mammalian Evolution* 17.3, pp. 223–225.

Fay, J. C., G. J. Wyckoff, and C.-I. Wu (2001). "Positive and negative selection on the human genome." *Genetics* 158.3, pp. 1227–1234.

FlyBase Consortium (2003). "The FlyBase database of the *Drosophila* genome projects and community literature." *Nucleic Acids Research* 31.1, pp. 172–175.

Ford, E. B. (1971). *Ecological genetics.* Chapman and Hall, p. 410. ISBN: 0412103206.

Frise, E., A. S. Hammonds, and S. E. Celniker (2010). "Systematic image-driven analysis of the spatial *Drosophila* embryonic expression landscape." *Molecular Systems Biology* 6.1, p. 345.

Futschik, M. E. and B. Carlisle (2005). "Noise-robust soft clustering of gene expression time-course data." *Journal of Bioinformatics and Computational Biology* 3.4, pp. 965–988.

Futschik, M. (2015). *Mfuzz: soft clustering of time series gene expression data*.

Gallo, S. M., D. T. Gerrard, D. Miner, M. Simich, B. Des Soye, C. M. Bergman, and M. S. Halfon (2011). "REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*." *Nucleic Acids Research* 39.Database, pp. D118–D123.

Galtier, N. (2016). "Adaptive protein evolution in animals and the effective population size hypothesis." *PLoS Genetics* 12.1, e1005774.

García-Dorado, A. (2012). "Understanding and predicting the fitness decline of shrunk populations: inbreeding, purging, mutation, and standard selection." *Genetics* 190.4, pp. 1461–1476.

Gelbart, W. M. and D. B. Emmert (2013). *FlyBase high throughput expression pattern data*. FlyBase Analysis.

Gellon, G. and W. McGinnis (1998). "Shaping animal body plans in development and evolution by modulation of Hox expression patterns." *BioEssays* 20.2, pp. 116–125.

Gerrish, P. J. and R. E. Lenski (1998). "The fate of competing beneficial mutations in an asexual population." *Genetica* 102-103.1-6, pp. 127–144.

Gerstein, M. B. et al. (2014). "Comparative analysis of the transcriptome across distant species." *Nature* 512.7515, pp. 445–448.

Gilbert, S. F. (2003). "The morphogenesis of evolutionary developmental biology." *The International Journal of Developmental Biology* 47.7-8, pp. 467–477.

Gilbert, S. F. (2014). *Developmental biology*. Sinauer Associates, p. 719. ISBN: 0878939784.

Gillespie, J. H. (2000a). "Genetic drift in an infinite population. The pseudohitchhiking model." *Genetics* 155.2, pp. 909–919.

Gillespie, J. H. (2000b). "The neutral theory in an infinite population." *Gene* 261.1, pp. 11–18.

Gillespie, J. H. (2001). "Is the population size of a species relevant to its evolution?" *Evolution* 55.11, pp. 2161–2169.

Gillespie, J. H. (2004). *Population genetics: a concise guide*. Johns Hopkins University Press, p. 214. ISBN: 0801880084.

Gillespie, J. H. (1991). *The causes of molecular evolution*. Oxford University Press, p. 336. ISBN: 0195092716.

Gingold, H. and Y. Pilpel (2011). "Determinants of translation efficiency and accuracy." *Molecular Systems Biology* 7, p. 481.

Giot, L. et al. (2003). "A protein interaction map of *Drosophila melanogaster*." *Science* 302.5651, pp. 1727–1736.

Goffeau, A. et al. (1996). "Life with 6000 genes." *Science* 274.5287, pp. 546–567.

Gossmann, T. I., P. D. Keightley, and A. Eyre-Walker (2012). "The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes." *Genome Biology and Evolution* 4.5, pp. 658–667.

Gossmann, T. I., D. Saleh, M. W. Schmid, M. A. Spence, and K. J. Schmid (2016). "Transcriptomes of plant gametophytes have a higher proportion of rapidly evolving and young genes than sporophytes." *Molecular Biology and Evolution* 33.7, pp. 1669–1678.

Gould, S. J. (1977). *Ontogeny and phylogeny*. Ed. by Harvard University Press. The Belknap Press of Harvard University Press, p. 501. ISBN: 0674639405.

Gramates, L. S. et al. (2017). "FlyBase at 25: looking to the future." *Nucleic Acids Research* 45.D1, pp. D663–D671.

Graveley, B. R. et al. (2011). "The developmental transcriptome of *Drosophila melanogaster*." *Nature* 471.7339, pp. 473–479.

Greenland, S. (1982). "Interpretation and estimation of summary ratios under heterogeneity." *Statistics in Medicine* 1.3, pp. 217–227.

Grenier, J. K., J. R. Arguello, M. C. Moreira, S. Gottipati, J. Mohammed, S. R. Hackett, R. Boughton, A. J. Greenberg, and A. G. Clark (2015). "Global diversity lines - a five-continent reference panel of sequenced *Drosophila melanogaster* strains." *G3* 5.4, pp. 593–603.

Gubb, D. (1986). "Intron-delay and the precision of expression of homoeotic gene products in *Drosophila*." *Developmental Genetics* 7.3, pp. 119–131.

Guillén, Y., S. Casillas, and A. Ruiz (2018). "Genome-wide patterns of sequence divergence of protein-coding genes between *Drosophila buzzatii* and *D. mojavensis*." *Journal of Heredity*.

Guo, Y.-L. (2013). "Gene family evolution in green plants with emphasis on the origination and evolution of A*rabidopsis thaliana* genes." *The Plant Journal* 73.6, pp. 941–951.

Gómez-Graciani, R. (2018). "Improving the interoperability between InvFEST and PopHuman." Master's thesis. Universitat Autònoma de Barcelona.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto (2005). "Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations." *Genome Research* 15.6, pp. 790–799.

Haddrill, P. R., D. L. Halligan, D. Tomaras, and B. Charlesworth (2007). "Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over." en. *Genome Biology* 8.2, R18.

Haeckel, E. (1879). *The evolution of man: a popular exposition of the principal points of human ontogeny and phylogeny.* Appleton, p. 536.

Haerty, W. and B. Golding (2009). "Similar selective factors affect both between-gene and between-exon divergence in *Drosophila*." *Molecular Biology and Evolution* 26.4, pp. 859–866.

Haerty, W. et al. (2007). "Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*." *Genetics* 177.3, pp. 1321–1335.

Hahn, M. W. (2018). *Molecular population genetics*. Sinauer Associates, p. 334. ISBN: 9780878939657.

Hales, K. G., C. A. Korey, A. M. Larracuente, and D. M. Roberts (2015). "Genetics on the fly: a primer on the Drosophila model system." *Genetics* 201.3, pp. 815–842.

Haller, B. C. and P. W. Messer (2017). "SLiM 2: flexible, interactive forward genetic simulations." *Molecular Biology and Evolution* 34.1, pp. 230–240.

Halligan, D. L. and P. D. Keightley (2006). "Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison." *Genome Research* 16.7, pp. 875–884.

Halligan, D. L., F. Oliver, A. Eyre-Walker, B. Harr, and P. D. Keightley (2010). "Evidence for pervasive adaptive protein evolution in wild mice." *PLoS Genetics* 6.1, e1000825.

Halligan, D. L., A. Kousathanas, R. W. Ness, B. Harr, L. Eöry, T. M. Keane, D. J. Adams, and P. D. Keightley (2013). "Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents." *PLoS Genetics* 9.12, e1003995.

Harris, H. (1966). "Enzyme polymorphisms in man." *Proceedings of the Royal Society of London. Series B, Biological sciences* 164.995, pp. 298–310.

Hartenstein, V. (1993). *Atlas of Drosophila development*. Cold Spring Harbor Laboratory Press. ISBN: 9780879694722.

Hashimshony, T., F. Wagner, N. Sher, and I. Yanai (2012). "CEL-seq: single-cell RNA-seq by multiplexed linear amplification." *Cell Reports* 2.3, pp. 666–673.

Hashimshony, T., M. Feder, M. Levin, B. K. Hall, and I. Yanai (2015). "Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer." *Nature* 519.7542, pp. 219–222.

Hastings, K. E. (1996). "Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families." *Journal of Molecular Evolution* 42.6, pp. 631–640.

Hebenstreit, D., M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, and S. A. Teichmann (2011). "RNA sequencing reveals two major classes of gene expression levels in metazoan cells." *Molecular Systems Biology* 7.1, p. 497.

Hernandez, R. D., S. H. Williamson, and C. D. Bustamante (2007). "Context dependence, ancestral misidentification, and spurious signatures of natural selection." *Molecular Biology and Evolution* 24.8, pp. 1792–1800.

Hershberg, R. and D. A. Petrov (2008). "Selection on codon bias." *Annual Review of Genetics* 42.1, pp. 287–299.

Hervas, S., E. Sanz, S. Casillas, J. E. Pool, and A. Barbadilla (2017). "PopFly: the *Drosophila* population genomics browser." *Bioinformatics* 33.17, pp. 2779–2780.

Heyn, P., M. Kircher, A. Dahl, J. Kelso, P. Tomancak, A. T. Kalinka, and K. M. Neugebauer (2014). "The earliest transcribed zygotic genes are short, newly evolved, and different across species." *Cell Reports* 6.2, pp. 285–292.

Heyn, P., A. T. Kalinka, P. Tomancak, and K. M. Neugebauer (2015). "Introns and gene expression: cellular constraints, transcriptional regulation, and evolutionary consequences." *BioEssays* 37.2, pp. 148–154.

Hill, W. G. and A. Robertson (1966). "The effect of linkage on limits to artificial selection." *Genetical Research* 8.3, pp. 269–294.

Hirsh, A. E. and H. B. Fraser (2001). "Protein dispensability and rate of evolution." *Nature* 411.6841, pp. 1046–1049.

Hirsh, A. E. and H. B. Fraser (2003). "Genomic function (communication arising): rate of evolution and gene dispensability." *Nature* 421.6922, pp. 497–498.

Hodgins, K. A., S. Yeaman, K. A. Nurkowski, L. H. Rieseberg, and S. N. Aitken (2016). "Expression divergence is correlated with sequence evolution but not positive selection in conifers." *Molecular Biology and Evolution* 33.6, pp. 1502–1516.

Holloway, A. K., D. J. Begun, A. Siepel, and K. S. Pollard (2008). "Accelerated sequence divergence of conserved genomic elements in *Drosophila melanogaster*." *Genome Research* 18.10, pp. 1592–1601.

Hu, T. T., M. B. Eisen, K. R. Thornton, and P. Andolfatto (2013). "A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence." *Genome Research* 23.1, pp. 89–98.

Huang, W. et al. (2014). "Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines." *Genome Research* 24.7, pp. 1193–1208.

Hurst, L. D. and N. G. Smith (1999). "Do essential genes evolve slowly?" *Current Biology* 9.14, pp. 747–750.

Irie, N. and S. Kuratani (2011). "Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis." *Nature Communications* 2, p. 248.

Irie, N. and S. Kuratani (2014). "The developmental hourglass model: a predictor of the basic body plan?" *Development* 141.24, pp. 4649–4655.

Irie, N. and A. Sehara-Fujisawa (2007). "The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information." *BMC Biology* 5, p. 1.

Jiggins, F. M. and K. W. Kim (2007). "A screen for immunity genes evolving under positive selection in *Drosophila*." *Journal of Evolutionary Biology* 20.3, pp. 965–970.

Jukes, T. H. and C. R. Cantor (1969). "Evolution of protein molecules." In: *Mammalian protein metabolism*. Academic Press. Chap. Evolution, pp. 21–32. ISBN: 9781483232119.

Kalinka, A. T. and P. Tomancak (2012). "The evolution of early animal embryos: conservation or divergence?" *Trends in Ecology & Evolution* 27.7, pp. 385–393.

Kalinka, A. T., K. M. Varga, D. T. Gerrard, S. Preibisch, D. L. Corcoran, J. Jarrells, U. Ohler, C. M. Bergman, and P. Tomancak (2010). "Gene expression divergence recapitulates the developmental hourglass model." *Nature* 468.7325, pp. 811–814.

Karaiskos, N., P. Wahle, J. Alles, A. Boltengagen, S. Ayoub, C. Kipar, C. Kocks, N. Rajewsky, and R. P. Zinzen (2017). "The *Drosophila* embryo at single-cell transcriptome resolution." *Science* 358.6360, pp. 194–199.

Karasov, T., P. W. Messer, and D. A. Petrov (2010). "Evidence that adaptation in *Drosophila* is not limited by mutation at single sites." *PLoS Genetics* 6.6, e1000924–e1000924.

Keightley, P. D. and A. Eyre-Walker (2007). "Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies." *Genetics* 177.4, pp. 2251–2261.

Keightley, P. D. and A. Eyre-Walker (2010). "What can we learn about the distribution of fitness effects of new mutations from DNA sequence data?" *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365.1544, pp. 1187–1193.

Keightley, P. D. and A. Eyre-Walker (2012). "Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small." *Journal of Molecular Evolution* 74.1-2, pp. 61–68.

Keightley, P. D. and B. C. Jackson (2018). "Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site." *Genetics* 209.3, pp. 897–906.

Kennison, J. A. (1993). "Transcriptional activation of *Drosphila* homeotic genes form distant regulatory elements." *Trends in Genetics* 9.3, pp. 75–79.

Kim, B. Y., C. D. Huber, and K. E. Lohmueller (2017). "Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples." *Genetics* 206.1, pp. 345–361.

Kimura, M. (1968). "Evolutionary rate at the molecular level." *Nature* 217.5129, pp. 624–626.

Kimura, M. (1977). "Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution." *Nature* 267.5608, pp. 275–276.

King, R. C. (1970). *Ovarian development in Drosophila melanogaster*. Academic Press, p. 227. ISBN: 9780124081505.

Kousathanas, A., F. Oliver, D. L. Halligan, and P. D. Keightley (2011). "Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice." *Molecular Biology and Evolution* 28.3, pp. 1183–1191.

Kreitman, M. (1983). "Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*." *Nature* 304.5925, pp. 412–417.

Kryuchkova-Mostacci, N. and M. Robinson-Rechavi (2017). "A benchmark of gene expression tissue-specificity metrics." *Briefings in Bioinformatics* 18.2, pp. 205–214.

Kumar, S. et al. (2011). "FlyExpress: visual mining of spatiotemporal patterns for genes and publications in *Drosophila* embryogenesis." *Bioinformatics* 27.23, pp. 3319–3320.

Kumar, S., G. Stecher, and K. Tamura (2016). "MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets." *Molecular Biology and Evolution* 33.7, pp. 1870–1874.

Kumar, S., C. Konikoff, M. Sanderford, L. Liu, S. Newfeld, J. Ye, and R. J. Kulathinal (2017). "FlyExpress 7: an integrated discovery platform to study coexpressed genes using in situ hybridization images in *Drosophila*." *G3* 7.8, pp. 2791–2797.

Künstner, A., B. Nabholz, and H. Ellegren (2011). "Significant selective constraint at 4-fold degenerate sites in the avian genome and its

consequence for detection of positive selection." *Genome Biology and Evolution* 3, pp. 1381–1389.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas, and M. Ashburner (1988). "Historical biogeography of the *Drosophila melanogaster* species subgroup." In: *Evolutionary biology*. Springer, pp. 159–225. ISBN: 9781461282518.

Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley, and J. E. Pool (2015). "The *Drosophila* Genome Nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population." *Genetics* 199.4, pp. 1229–1241.

Lack, J. B., J. D. Lange, A. D. Tang, R. B. Corbett-Detig, and J. E. Pool (2016). "A thousand fly genomes: an expanded *Drosophila* Genome Nexus." *Molecular Biology and Evolution* 33.12, pp. 3308–3313.

Lanfear, R., H. Kokko, and A. Eyre-Walker (2014). "Population size and the rate of evolution." *Trends in Ecology & Evolution* 29.1, pp. 33–41.

Langley, C. H. et al. (2012). "Genomic variation in natural populations of *Drosophila melanogaster*." *Genetics* 192.2, pp. 533–598.

Larracuente, A. M. et al. (2008). "Evolution of protein-coding genes in *Drosophila*." *Trends in Genetics* 24.3, pp. 114–123.

Lawrie, D. S., P. W. Messer, R. Hershberg, and D. A. Petrov (2013). "Strong purifying selection at synonymous sites in *D. melanogaster*." *PLoS Genetics* 9.5, e1003527–e1003527.

Leader, D. P., S. A. Krause, A. Pandit, S. A. Davies, and J. A. T. Dow (2018). "FlyAtlas 2: a new version of the *Drosophila melanogaster* expression atlas with RNA-Seq, miRNA-Seq and sex-specific data." *Nucleic Acids Research* 46.D1, pp. D809–D815.

Lécuyer, E., H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T. R. Hughes, P. Tomancak, and H. M. Krause (2007). "Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function." *Cell* 131.1, pp. 174–187.

Lefevre, G. and U. B. Jonsson (1962). "Sperm transfer, storage, displacement, and utilization in *Drosophila melanogaster*." *Genetics* 47.12, pp. 1719–1736.

Lemos, B., B. R. Bettencourt, C. D. Meiklejohn, and D. L. Hartl (2005). "Evolution of proteins and gene expression levels are coupled in textitDrosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions." *Molecular Biology and Evolution* 22.5, pp. 1345–1354.

Leung, W. et al. (2010). "Evolution of a distinct genomic domain in *Drosophila*: comparative analysis of the dot chromosome

in *Drosophila melanogaster* and *Drosophila virilis*." *Genetics* 185.4, pp. 1519–1534.

Levin, M., T. Hashimshony, F. Wagner, and I. Yanai (2012). "Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo." *Developmental Cell* 22.5, pp. 1101–1108.

Levin, M. et al. (2016). "The mid-developmental transition and the evolution of animal body plans." *Nature* 531.7596, pp. 637–641.

Lewontin, R. C. (1991). "Twenty-five years ago in genetics: electrophoresis in the development of evolutionary genetics: milestone or millstone?" *Genetics* 128.4.

Lewontin, R. C. (2000). "The problems of population genetics." In: *Evolutionary genetics: from molecules to morphology*. Cambridge University Press, p. 702. ISBN: 0521571235.

Lewontin, R. C. and J. L. Hubby (1966). "A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*." *Genetics* 54.2, pp. 595–609.

Lewontin, R. C. (1974). *The genetic basis of evolutionary change*. Columbia University Press, p. 346. ISBN: 0231083181.

Li, H. and W. Stephan (2006). "Inferring the demographic history and rate of adaptive substitution in *Drosophila*." *PLoS Genetics* 2.10, e166.

Li, W. H. (1987). "Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons." *Journal of Molecular Evolution* 24.4, pp. 337–345.

Loewe, L. (2009). "A framework for evolutionary systems biology." *BMC Systems Biology* 3.1, p. 27.

Lynch, M. (2006). "The origins of eukaryotic gene structure." *Molecular Biology and Evolution* 23.2, pp. 450–468.

Lynch, M. (2007). *The origins of genome architecture*. Sinauer Associates, p. 494. ISBN: 0878934847.

Lyne, R. et al. (2007). "FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics." *Genome Biology* 8.7, R129.

Mackay, T. F. C. et al. (2012). "The *Drosophila melanogaster* Genetic Reference Panel." *Nature* 482.7384, pp. 173–178.

Marais, G., T. Domazet-Loso, D. Tautz, and B. Charlesworth (2004). "Correlated evolution of synonymous and nonsynonymous sites in *Drosophila*." *Journal of Molecular Evolution* 59.6, pp. 771–779.

Marais, G., P. Nouvellet, P. D. Keightley, and B. Charlesworth (2005). "Intron size and exon evolution in *Drosophila*." *Genetics* 170.1, pp. 481–485.

Markow, T. A. and P. M. O'Grady (2006). *Drosophila: a guide to species identification and use*. Elsevier, p. 259. ISBN: 0080454097.

Markow, T. A. and P. M. O'Grady (2007). "*Drosophila* biology in the genomic age." *Genetics* 177.3, pp. 1269–1276.

Marra, N. J., A. Romero, and J. A. DeWoody (2014). "Natural selection and the genetic basis of osmoregulation in heteromyid rodents as revealed by RNA-seq." *Molecular Ecology* 23.11, pp. 2699–2711.

Martin, A. and V. Orgogozo (2013). "The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation." *Evolution* 67.5, pp. 1235–1250.

McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the *Adh* locus in *Drosophila*." *Nature* 351.6328, pp. 652–654.

McQuilton, P., S. E. St. Pierre, J. Thurmond, and FlyBase Consortium (2012). "FlyBase 101 - the basics of navigating FlyBase." *Nucleic Acids Research* 40.D1, pp. D706–D714.

Meiklejohn, C. D., J. Parsch, J. M. Ranz, and D. L. Hartl (2003). "Rapid evolution of male-biased gene expression in *Drosophila*." *Proceedings of the National Academy of Sciences of the United States of America* 100.17, pp. 9894–9899.

Mensch, J., F. Serra, N. J. Lavagnino, H. Dopazo, and E. Hasson (2013). "Positive selection in nucleoporins challenges constraints on early expressed genes in *Drosophila* development." *Genome Biology and Evolution* 5.11, pp. 2231–2241.

Messer, P. W. and D. a. Petrov (2013). "Frequent adaptation and the McDonald-Kreitman test." *Proceedings of the National Academy of Sciences of the United States of America* 110.21, pp. 8615–8620.

Mi, H., X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P. D. Thomas (2017). "PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements." *Nucleic Acids Research* 45.D1, pp. D183–D189.

Morgan, T. H. (1910). "Sex limited inheritance in *Drosophila*." *Science* 32.812, pp. 120–122.

Morisalo, D. and K. V. Anderson (1995). "Signaling pathways that establish the dorsal-ventral pattern of the Drosophila embryo." *Annual Review of Genetics* 29.1, pp. 371–399.

Muller, H. J. and W. D. Kaplan (1966). "The dosage compensation of *Drosophila* and mammals as showing the accuracy of the normal type." *Genetical Research* 8.1, pp. 41–59.

Murga-Moreno, J., M. Coronado-Zamora, A. Bodelón, A. Barbadilla, and S. Casillas (2018). "PopHumanScan: the collaborative catalog of human adaptation." *Nucleic Acid Research*.

Nielsen, R. and M. Slatkin (2013). *An introduction to population genetics: theory and applications*. Sinauer Associates, p. 298. ISBN: 9781605351537.

O'Farrell, P. H., J. Stumpff, and T. T. Su (2004). "Embryonic cleavage cycles: how is a mouse like a fly?" *Current Biology* 14.1, pp. 35–45.

Obbard, D. J., J. J. Welch, K.-W. Kim, and F. M. Jiggins (2009). "Quantifying adaptive evolution in the *Drosophila* immune system." *PLoS Genetics* 5.10, e1000698.

Ohta, T. (1973). "Slightly deleterious mutant substitutions in evolution." *Nature* 246.5428, pp. 96–98.

Pagès, H., M. Carlson, S. Falcon, and N. Li (2017). "AnnotationDbi: Annotation Database Interface." R package version 1.38.2.

Pál, C., B. Papp, and M. J. Lercher (2006). "An integrated view of protein evolution." *Nature reviews. Genetics* 7.5, pp. 337–348.

Pal, C., B. Papp, L. D. Hurst, C Pál, B. Papp, and L. D. Hurst (2001). "Highly expressed genes in yeast evolve slowly." *Genetics* 158.2, pp. 927–931.

Pantalacci, S. and M. Sémon (2015). "Transcriptomics of developing embryos and organs: A raising tool for evo-devo." *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 324.4, pp. 363–371.

Parsch, J., S. Novozhilov, S. S. Saminadin-Peter, K. M. Wong, and P. Andolfatto (2010). "On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*." *Molecular Biology and Evolution* 27.6, pp. 1226–1234.

Peck, J. R. (1994). "A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex." *Genetics* 137.2, pp. 597–606.

Peden, J. F. (1999). "Analysis of codon usage." PhD thesis. University of Nottingham.

Petit, N., S. Casillas, A. Ruiz, and A. Barbadilla (2007). "Protein polymorphism is negatively correlated with conservation of intronic sequences and complexity of expression patterns in *Drosophila melanogaster*." *Journal of Molecular Evolution* 64.5, pp. 511–518.

Phung, T. N., C. D. Huber, and K. E. Lohmueller (2016). "Determining the effect of natural selection on linked neutral divergence across species." *PLoS Genetics* 12.8, e1006199.

Piasecka, B., P. Lichocki, S. Moretti, S. Bergmann, and M. Robinson-Rechavi (2013). "The hourglass and the early conservation models-co-existing patterns of developmental constraints in vertebrates." *PLoS Genetics* 9.4, e1003476.

Plotkin, J. B. and H. B. Fraser (2007). "Assessing the determinants of evolutionary rates in the presence of noise." *Molecular Biology and Evolution* 24.5, pp. 1113–1121.

Poe, S. and M. H. Wake (2004). "Quantitative tests of general models for the evolution of development." *The American Naturalist* 164.3, pp. 415–422.

Pool, J. E. et al. (2012). "Population Genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture." *PLoS Genetics* 8.12, e1003080.

Popescu, C. E., T. Borza, J. P. Bielawski, and R. W. Lee (2006). "Evolutionary rates and expression level in *Chlamydomonas*." *Genetics* 172.3, pp. 1567–76.

Powell, J. R. (1997). *Progress and prospects in evolutionary biology: the Drosophila model*. Oxford University Press, p. 562. ISBN: 019536032X.

Presgraves, D. C. (2005). "Recombination enhances protein adaptation in *Drosophila melanogaster*." *Current Biology* 15.18, pp. 1651–1656.

Pröschel, M., Z. Zhang, and J. Parsch (2006). "Widespread adaptive evolution of *Drosophila* genes with sex-biased expression." *Genetics* 174.2, pp. 893–900.

Quax, T. E., N. J. Claassens, D. Söll, and J. van der Oost (2015). "Codon bias as a means to fine-tune gene expression." *Molecular Cell* 59.2, pp. 149–161.

Quint, M., H.-G. Drost, A. Gabel, K. K. Ullrich, M. Bönn, and I. Grosse (2012). "A transcriptomic hourglass in plant embryogenesis." *Nature* 490.7418, pp. 98–101.

R Core Team (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing.

Raff, R. A. (1996). *The shape of life: genes, development, and the evolution of animal form*. University of Chicago Press, p. 520. ISBN: 0226702669.

Raff, R. A. (2000). "Evo-devo: the evolution of a new discipline." *Nature Reviews Genetics* 1.1, pp. 74–79.

Ràmia, M. (2015). "Visualization, description and analysis of the genome variation of a natural population of *Drosophila melanogaster*." PhD thesis. Universitat Autònoma de Barcelona.

Ràmia, M., P. Librado, S. Casillas, J. Rozas, and A. Barbadilla (2012). "PopDrowser: the population *Drosophila* browser." *Bioinformatics* 28.4, pp. 595–596.

Rand, D. M. and L. M. Kann (1996). "Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans." *Molecular Biology and Evolution* 13.6, pp. 735–748.

Rao, Y. S., Z. F. Wang, X. W. Chai, G. Z. Wu, M. Zhou, Q. H. Nie, and X. Q. Zhang (2010). "Selection for the compactness of highly expressed genes in *Gallus gallus*." *Biology Direct* 5.1, p. 35.

Ratheesh, A., V. Belyaeva, and D. E. Siekhaus (2015). "*Drosophila* immune cell migration and adhesion during embryonic development and larval immune responses." *Current Opinion in Cell Biology* 36, pp. 71–79.

Richardson, M. K., J Hanken, M. L. Gooneratne, C Pieau, A Raynaud, L Selwood, and G. M. Wright (1997). "There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development." *Anatomy and Embryology* 196.2, pp. 91–106.

Richardson, M. K. (1995). "Heterochrony and the phylotypic period." *Developmental Biology* 172.2, pp. 412–421.

Riedl, R. (1978). *Order in living organisms: a systems analysis of evolution*. Wiley. ISBN: 0471996351.

Rinker, D. C., X. Zhou, R. Pitts, A. Rokas, and L. J. Zwiebel (2013). "Antennal transcriptome profiles of anopheline mosquitoes reveal human host olfactory specialization in Anopheles gambiae." *BMC Genomics* 14.1, p. 749.

Rocha, E. P. C. (2006). "The quest for the universals of protein evolution." *Trends in Genetics* 22.8, pp. 412–416.

Rocha, E. P. C. and A. Danchin (2004). "An analysis of determinants of amino acids substitution rates in bacterial proteins." *Molecular Biology and Evolution* 21.1, pp. 108–116.

Rogers, R. L., J. M. Cridland, L. Shao, T. T. Hu, P. Andolfatto, and K. R. Thornton (2014). "Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*." *Molecular Biology and Evolution* 31.7, pp. 1750–1766.

Roux, J. and M. Robinson-Rechavi (2008). "Developmental constraints on vertebrate genome evolution." *PLoS Genetics* 4.12, e1000311.

Roux, J., M. Rosikiewicz, and M. Robinson-Rechavi (2015). "What to compare and how: comparative transcriptomics for evo-devo." *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 324.4, pp. 372–382.

Rubin, G. M. (1996). "Around the genomes: the *Drosophila* genome project." *Genome Research* 6.2, pp. 71–79.

Salathe, M., M. Ackermann, S. Bonhoeffer, M. Salathé, M. Ackermann, and S. Bonhoeffer (2006). "The effect of multifunctionality on the rate of evolution in yeast." *Molecular Biology and Evolution* 23.4, pp. 721–722.

Salazar-Ciudad, I., J. Jernvall, and S. A. Newman (2003). "Mechanisms of pattern formation in development and evolution." *Development* 130.10, pp. 2027–2037.

Salvador-Martínez, I. and I. Salazar-Ciudad (2015). "How complexity increases in development: an analysis of the spatial-temporal dynamics of 1218 genes in *Drosophila melanogaster*." *Developmental Biology* 405.2, pp. 328–339.

Salvador-Martínez, I. and I. Salazar-Ciudad (2017). "How complexity increases in development: An analysis of the spatial-temporal dynamics of gene expression in *Ciona intestinalis*." *Mechanisms of Development* 144.Pt B, pp. 113–124.

Salvador-Martínez, I., M. Coronado-Zamora, D. Castellano, A. Barbadilla, and I. Salazar-Ciudad (2018). "Mapping selection within *Drosophila melanogaster* embryo's anatomy." *Molecular Biology and Evolution* 35.1, pp. 66–79.

Sander, K. (1983). "The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis." In: *Development and evolution*. Cambridge University Press, pp. 137–159.

Sawyer, S. A., R. J. Kulathinal, C. D. Bustamante, and D. L. Hartl (2003). "Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection." *Journal of Molecular Evolution* 57 Suppl 1, pp. 154–164.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* 270.5235, pp. 467–470.

Schlenke, T. A. and D. J. Begun (2003). "Natural selection drives *Drosophila* immune system evolution." *Genetics* 164.4, pp. 1471–1480.

Schmidt, P. S., L. Matzkin, M. Ippolito, and W. F. Eanes (2005). "Geographic variation in diapause incidence, life-history traits, and climatic adaptation in *Drosophila melanogaster*." *Evolution* 59.8, pp. 1721–1732.

Shah, S. et al. (2018). "Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH." *Cell* 174.2, 363–376.e16.

Sham, P. C. and S. M. Purcell (2014). "Statistical power and significance testing in large-scale genetic studies." *Nature Reviews Genetics* 15.5, pp. 335–346.

Siefert, J. C., E. A. Clowdus, and C. L. Sansam (2015). "Cell cycle control in the early embryonic development of aquatic animal species." *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* 178, pp. 8–15.

Signor, S. A., F. N. New, and S. Nuzhdin (2018). "A large panel of *Drosophila simulans* reveals an abundance of common variants." *Genome Biology and Evolution* 10.1, pp. 189–206.

Simpson, E. H (1951). "The interpretation of interaction in contingency tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 13.2, pp. 238–241.

Singh, R. S. and L. R. Rhomberg (1987). "A comprehensive study of genic cariation in natural populations of *Drosophila melanogaster*. I. Estimates of gene flow from rare alleles." *Genetics* 115.2, pp. 313–322.

Slack, J. M. W., P. W. H. Holland, and C. F. Graham (1993). "The zootype and the phylotypic stage." *Nature* 361.6412, pp. 490–492.

Smith, J. M. and J. Haigh (1974). "The hitch-hiking effect of a favourable gene." *Genetical Research* 23.1, pp. 23–35.

Smith, N. G. C. and A. Eyre-Walker (2002). "Adaptive protein evolution in *Drosophila*." *Nature* 415.6875, pp. 1022–1024.

Sorek, R. and G. Ast (2003). "Intronic sequences flanking alternatively spliced exons are conserved between human and mouse." *Genome Research* 13.7, pp. 1631–1637.

Spitz, F. and E. E. M. Furlong (2012). "Transcription factors: from enhancer binding to developmental control." *Nature Reviews Genetics* 13.9, pp. 613–626.

St Pierre, S. E., L. Ponting, R. Stefancsik, and P. McQuilton (2014). "FlyBase 102 - advanced approaches to interrogating FlyBase." *Nucleic Acids Research* 42.Database issue, pp. D780–D788.

Stahl, P. L. et al. (2016). "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics." *Science* 353.6294, pp. 78–82.

Stanley, C. E. and R. J. Kulathinal (2016). "flyDIVaS: a comparative genomics resource for *Drosophila* divergence and selection." *G3* 6.8, pp. 2355–2363.

Stephan, W. and H. Li (2007). "The recent demographic and adaptive history of *Drosophila melanogaster*." *Heredity* 98.2, pp. 65–68.

Stocker, H. and P. Gallant (2008). "Getting Started: an overview on raising and handling *Drosophila*." In: *Methods in molecular biology*. Vol. 420, pp. 27–44. ISBN: 9781588298171.

Stoletzki, N. and A. Eyre-Walker (2011). "Estimation of the neutrality index." *Molecular Biology and Evolution* 28.1, pp. 63–70.

Stone, E. A. (2012). "Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines." *Genome Research* 22.5, pp. 966–974.

Subramanian, S. and S. Kumar (2004). "Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome." *Genetics* 168.1, pp. 373–381.

Swanson, W. J., Z. Yang, M. F. Wolfner, and C. F. Aquadro (2001). "Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals." *Proceedings of the National Academy of Sciences* 98.5, pp. 2509–2514.

Swanson, W. J. and V. D. Vacquier (2002). "The rapid evolution of reproductive proteins." *Nature Reviews Genetics* 3.2, pp. 137–144.

Swinburne, I. A. and P. A. Silver (2008). "Intron delays and transcriptional timing during development." *Developmental Cell* 14.3, pp. 324–330.

Takashima, Y., T. Ohtsuka, A. Gonzalez, H. Miyachi, and R. Kageyama (2011). "Intronic delay is essential for oscillatory expression in the segmentation clock." *Proceedings of the National Academy of Sciences* 108.8, pp. 3300–3305.

Tamura, K., S. Subramanian, and S. Kumar (2004). "Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks." *Molecular Biology and Evolution* 21.1, pp. 36–44.

Tarone, R. E. (1981). "On summary estimators of relative risk." *Journal of Chronic Diseases* 34.9-10, pp. 463–468.

Tataru, P., M. Mollion, S. Glémin, and T. Bataillon (2017). "Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data." *Genetics*.

Technau, U. and C. B. Scholz (2003). "Origin and evolution of endoderm and mesoderm." *The International Journal of Developmental Biology* 47.7-8, pp. 531–539.

Templeton, A. R. (1996). "Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates." *Genetics* 144.3, pp. 1263–1270.

Thomas, G. W. C. et al. (2018). "The genomic basis of arthropod diversity." *bioRxiv*, p. 382945.

Thomsen, S., S. Anders, S. C. Janga, W. Huber, and C. R. Alonso (2010). "Genome-wide analysis of mRNA decay patterns during early *Drosophila* development." *Genome Biology* 11.9, R93.

Thornton, K. and P. Andolfatto (2006). "Approximate bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*." *Genetics* 172.3, pp. 1607–1619.

Tomancak, P. et al. (2002). "Systematic determination of patterns of gene expression during *Drosophila* embryogenesis." *Genome Biology* 3.12, research0088.1.

Tomancak, P., B. P. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S. E. Celniker, and G. M. Rubin (2007). "Global analysis of patterns of gene expression during *Drosophila* embryogenesis." *Genome Biology* 8.7, R145.

Vining, K. H. and D. J. Mooney (2017). "Mechanical forces direct stem cell behaviour in development and regeneration." *Nature Reviews Molecular Cell Biology* 18.12, pp. 728–742.

Vitti, J. J., S. R. Grossman, and P. C. Sabeti (2013). "Detecting natural selection in genomic data." *Annual Review of Genetics* 47.1, pp. 97–120.

Wade, M. J., N. K. Priest, and T. E. Cruickshank (2009). "A theoretical overview of genetic maternal effects." In: *Maternal effects in mammals*. University of Chicago Press, pp. 38–63. ISBN: 9780226501208.

Wagner, A. (2008). "Neutralism and selectionism: a network-based reconciliation." *Nature Reviews Genetics* 9.12, pp. 965–974.

Wagner, G. P., K. Kin, and V. J. Lynch (2013). "A model based criterion for gene expression calls using RNA-seq data." *Theory in Biosciences* 132.3, pp. 159–164.

Walsh, B. and M. Lynch (2018). *Evolution and selection of quantitative traits*. Oxford University Press. ISBN: 0192566644.

Wang, Z., M. Gerstein, and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." *Nature Reviews Genetics* 10.1, pp. 57–63.

Wang, Z. et al. (2013). "The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan." *Nature Genetics* 45.6, pp. 701–706.

Watterson, G. A. (1975). "On the number of segregating sites in genetical models without recombination." *Theoretical Population Biology* 7.2, pp. 256–276.

Westfall, P. H. and S. S Young (1993). *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley, p. 340. ISBN: 0471557617.

Wilk, R., J. Hu, D. Blotsky, and H. M. Krause (2016). "Diverse and pervasive subcellular distributions for both coding and long noncoding RNAs." *Genes & Development* 30.5, pp. 594–609.

Williamson, R. J., E. B. Josephs, A. E. Platts, K. M. Hazzouri, A. Haudry, M. Blanchette, and S. I. Wright (2014). "Evidence for widespread positive and negative selection in coding and conserved noncod-

ing regions of *Capsella grandiflora*." *PLoS Genetics* 10.9. Ed. by M. W. Nachman, e1004622.

Wilson, A. C., S. S. Carlson, and T. J. White (1977). "Biochemical evolution." *Annual Review of Biochemistry* 46.1, pp. 573–639.

Wilt, F. H. and S. Hake (2004). *Principles of developmental biology*. W.W. Norton, p. 430. ISBN: 0393974308.

Wimsatt, W. C. (1986). *Developmental constraints, generative entrenchment, and the innate-acquired distinction*. Springer, pp. 185–208. ISBN: 9789024733422.

Wolf, J. B. W., A. Künstner, K. Nam, M. Jakobsson, and H. Ellegren (2009a). "Nonlinear dynamics of nonsynonymous (d_N) and synonymous (d_S) substitution rates affects inference of selection." *Genome Biology and Evolution* 1.0, pp. 308–319.

Wolf, Y. I., P. S. Novichkov, G. P. Karev, E. V. Koonin, and D. J. Lipman (2009b). "The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages." *Proceedings of the National Academy of Sciences of the United States of America* 106.18, pp. 7273–7280.

Wolpert, L. L. (1991). *The triumph of the embryo*. Oxford University Press, p. 211. ISBN: 0198542437.

Wotton, K. R., E. Jiménez-Guri, A. Crombach, D. Cicin-Sain, and J. Jaeger (2015). "High-resolution gene expression data from blastoderm embryos of the scuttle fly *Megaselia abdita*." *Scientific Data* 2, p. 150005.

Wray, G. A. (2000). "The evolution of embryonic patterning mechanisms in animals." *Seminars in Cell & Developmental Biology* 11.6, pp. 385–393.

Wray, G. A. (2002). "Do convergent developmental mechanisms underlie convergent phenotypes?" *Brain, Behavior and Evolution* 59.5-6, pp. 327–336.

Wright, S. I., C. B. K. Yau, M. Looseley, and B. C. Meyers (2004). "Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and textitArabidopsis lyrata." *Molecular Biology and Evolution* 21.9, pp. 1719–1726.

Wyckoff, G. J., W. Wang, and C. I. Wu (2000). "Rapid evolution of male reproductive genes in the descent of man." *Nature* 403.6767, pp. 304–309.

Yanai, I. (2018). "Development and evolution through the lens of global gene regulation." *Trends in Genetics* 34.1, pp. 11–20.

Yanai, I. et al. (2005). "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification." *Bioinformatics* 21.5, pp. 650–659.

Yanai, I., L. Peshkin, P. Jorgensen, and M. W. Kirschner (2011). "Mapping gene expression in two *Xenopus* species: evolutionary constraints and developmental flexibility." *Developmental Cell* 20.4, pp. 483–496.

Yang, Z. and J. P. Bielawski (2000). "Statistical methods for detecting molecular adaptation." *Trends in Ecology & Evolution* 15.12, pp. 496–503.

Yu, J., S. Pacifico, G. Liu, and R. L. Finley Jr. (2008). "DroID: the *Drosophila* Interactions Database, a comprehensive resource for annotated gene and protein interactions." *BMC Genomics* 9.1, p. 461.

Zhang, L. and W.-H. Li (2004). "Mammalian housekeeping genes evolve more slowly than tissue-specific genes." *Molecular Biology and Evolution* 21.2, pp. 236–239.

Zhang, Z. and J. Parsch (2005). "Positive correlation between evolutionary rate and recombination rate in *Drosophila* genes with male-biased expression." *Molecular Biology and Evolution* 22.10, pp. 1945–1947.

Zhou, J., H. Y. Kim, L. A. Davidson, S. Zhan, M. D. Schneider, F. J. DeMayo, and R. J. Schwartz (2009). "Actomyosin stiffens the vertebrate embryo during crucial stages of elongation and neural tube closure." *Development* 136.4, pp. 677–688.

Zuckerkandl, E. (1976). "Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins." *Journal of Molecular Evolution* 7.3, pp. 167–183.

Zuckerkandl, E. and L. Pauling (1965). "Evolutionary divergence and convergence in proteins." In: *Evolving genes and proteins*. Elsevier, pp. 97–166.

Chapter 6

APPENDIX

Supplementary figures

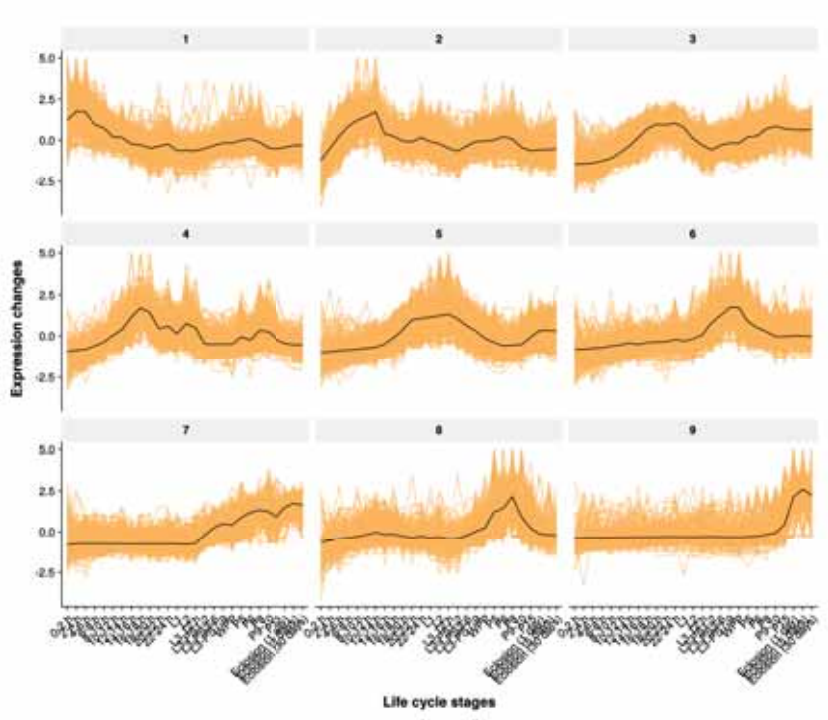


Figure A.1 Temporal profile of expression of the genes in each of the nine life cycle clusters. In yellow is represented the expression profile of the genes belonging to a cluster with a membership $\geqslant 0.8$.

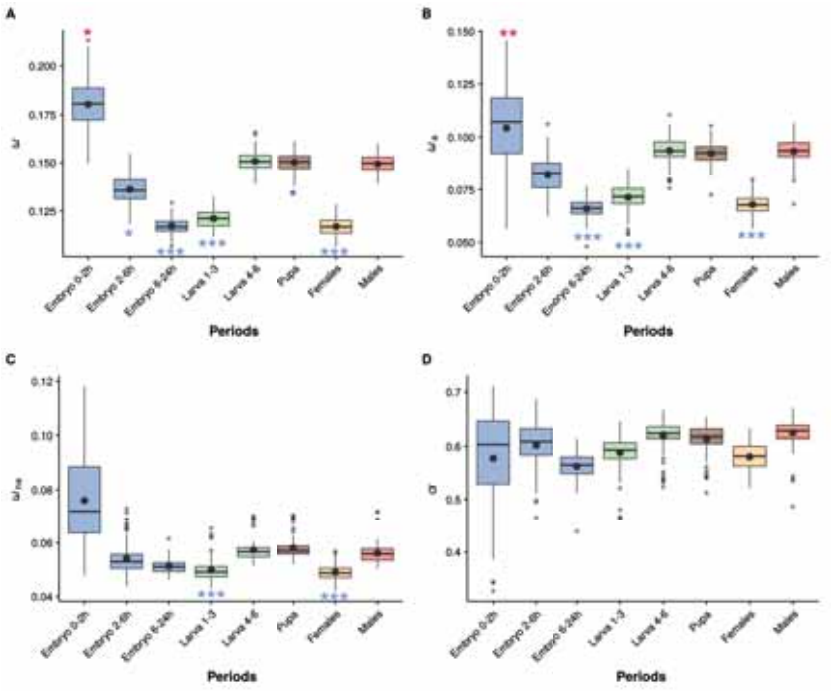


Figure A.2 ω, ω_a, ω_{na} and α of the different developmental periods estimated using DFE-alpha using the life cycle set genes for the null distribution. A. ω, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **B**. α, the proportion of base substitutions fixed by natural selection. **C**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **D**. ω, the rate of non-synonymous substitutions relative to the mutation rate. Each boxplot (100 bootstrap replicates per period) in a plot is calculated for a randomly drawn sample of the set of genes that belong in a period with replacement. The number of genes expressed in each period is: Embryo 0-2h: 333 genes; Embryo 2-6h: 886 genes; Embryo 6-24h: 2,086 genes; Larva begin: 1,877 genes; Larva end: 2,363 genes; Pupa: 2,708 genes; Females: 1,784 genes; Males: 2,563 genes. *P*-values in Table B.16.

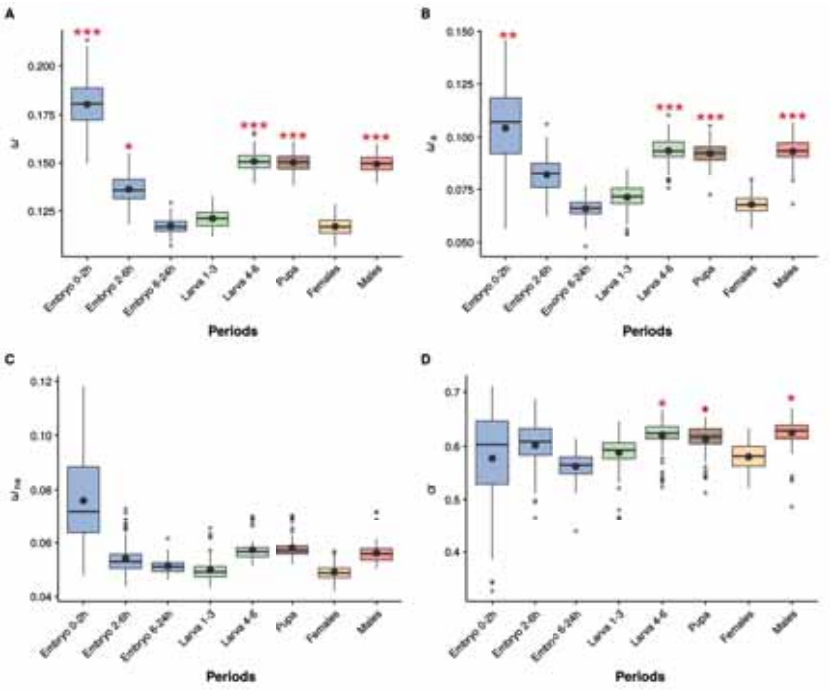


Figure A.3 ω, ω_a, ω_{na} and α of the different developmental periods estimated using DFE-alpha using the life cycle set genes for the null distribution. A. ω, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **B**. α, the proportion of base substitutions fixed by natural selection. **C**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **D**. ω, the rate of non-synonymous substitutions relative to the mutation rate. Each boxplot (100 bootstrap replicates per period) in a plot is calculated for a randomly drawn sample of the set of genes that belong in a period with replacement. The number of genes expressed in each period is: Embryo 0-2h: 333 genes; Embryo 2-6h: 886 genes; Embryo 6-24h: 2,086 genes; Larva begin: 1,877 genes; Larva end: 2,363 genes; Pupa: 2,708 genes; Females: 1,784 genes; Males: 2,563 genes. *P*-values in Table B.17.

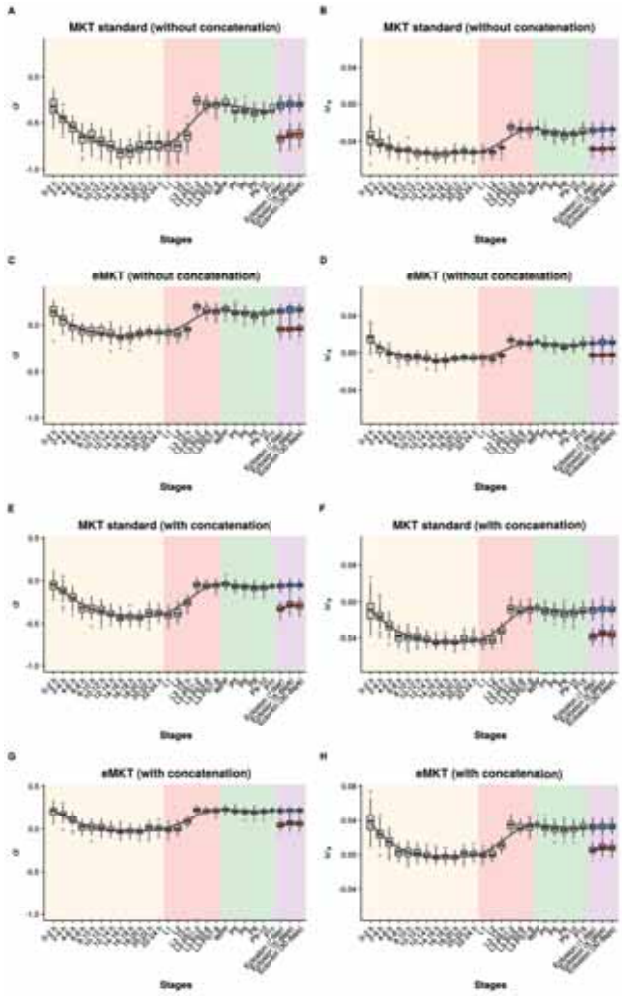


Figure A.4 The selection statistics follow a similar temporal pattern when measured with the eMKT and standard MKT method. **A**. α, estimated using the standard MKT, without concatenating genes. **B**. ω_a, estimated using the standard MKT method, without concatenating genes. **C**. α, estimated using the eMKT method, without concatenating genes. **D**. ω_a, estimated using the eMKT method, without concatenating genes. In Figures **A-D**, $\alpha$and ω_a were estimated individually for each gene expressed in each stage (number of genes analyzed in Table B.8). Each boxplot **A-D** represents the median estimated by the bootstrap method (100 times with replacement). **E**. α, estimated using the standard MKT, concatenating genes. **F**. ω_a, estimated using the standard MKT method, concatenating genes. **G**. α, estimated using the eMKT method, concatenating genes. **H**. ω_a, estimated using the eMKT method, concatenating genes. In this case, the same data as for Figure 3.9 was used. Colors and lines as in Figure 3.9.

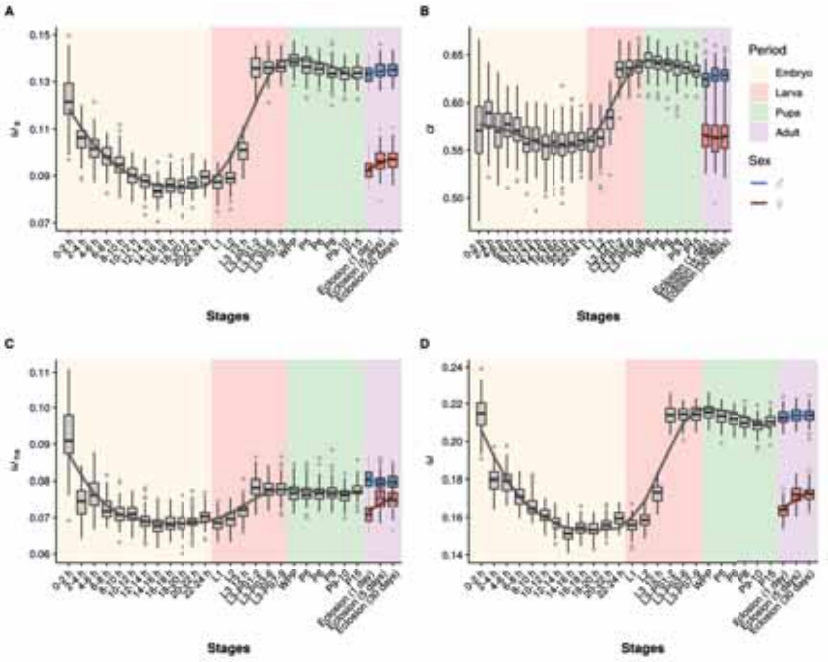


Figure A.5 ω_a, α, ω_{na} and ω over estimated using DFE-alpha developmental time when using 4-fold as a proxy for the mutation rate. **A**. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **B**. α, the proportion of base substitutions fixed by natural selection. **C**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **D**. ω, the rate of non-synonymous substitutions relative to the mutation rate. Colors and lines as in Figure 3.9. The number of analyzed genes is shown in Table B.1.

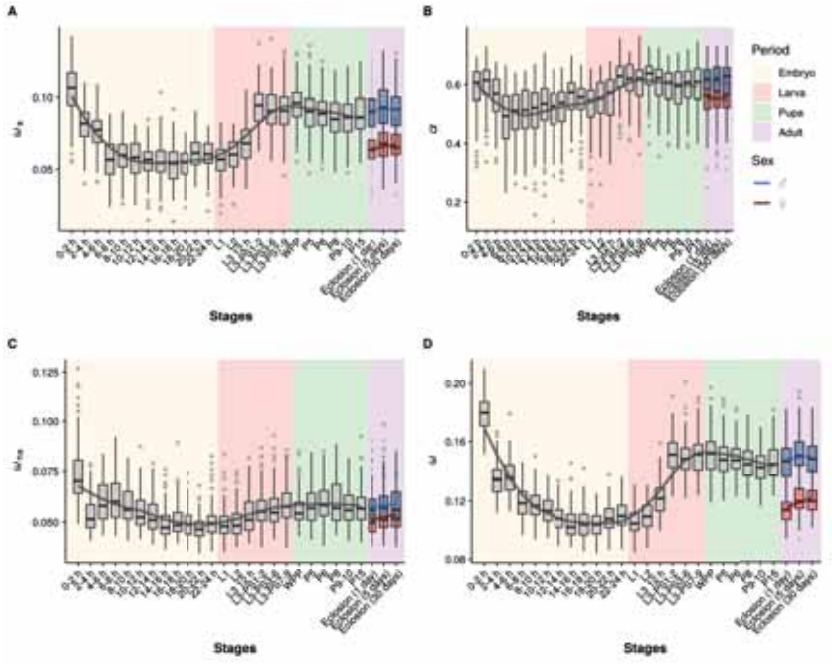


Figure A.6 ω_a, α, ω_{na} and ω estimated using DFE-alpha over developmental stages when resampling the same number of genes in each stage. A. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **B**. α, the proportion of base substitutions fixed by natural selection. **C**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **D**. ω, the rate of non-synonymous substitutions relative to the mutation rate. Colors and lines as in Figure 3.9. The number of analyzed genes is shown in Table B.1.

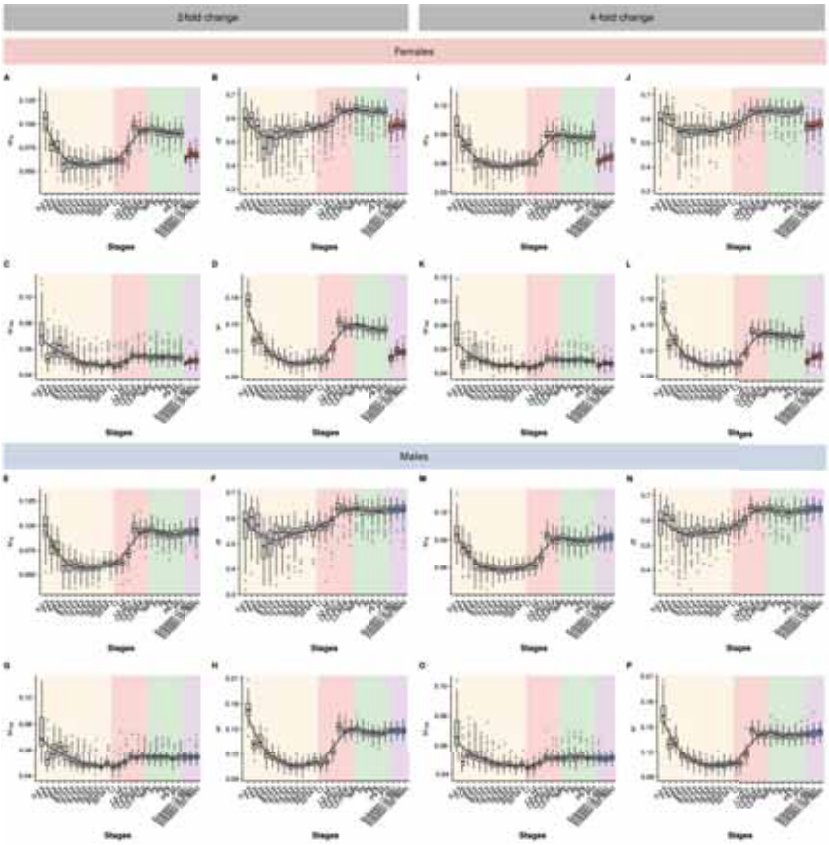


Figure A.7 ω_a, α, ω_{na} and ω estimated using DFE-alpha over developmental time when analyzing genes that have a maximal level of expression that is at least twice or four times than of its minimal expression for females and males. In this analysis, only genes that have a maximal level of expression (over all stages) that is at least twice (or four times) than of its minimal expression are considered. In all cases, the pattern resembles that of Figure 3.9. **A**. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **B**. α, the proportion of base substitutions fixed by natural selection. **C**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **D**. ω, the rate of non-synonymous substitutions relative to the mutation rate. **E**. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **F**. α, the proportion of base substitutions fixed by natural selection. **G**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **H**. ω, the rate of non-synonymous substitutions relative to the mutation rate. **I**. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **J**. α, the proportion of base substitutions fixed by natural selection. **K**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **L**. ω, the rate of non-synonymous substitutions relative to the mutation rate. **M**. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **N**. α, the proportion of base substitutions fixed by natural selection. **O**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **P**. ω, the rate of non-synonymous substitutions relative to the mutation rate. Colors and lines as in Figure 3.9. The number of analyzed genes is shown in Tables B.2-B.5.

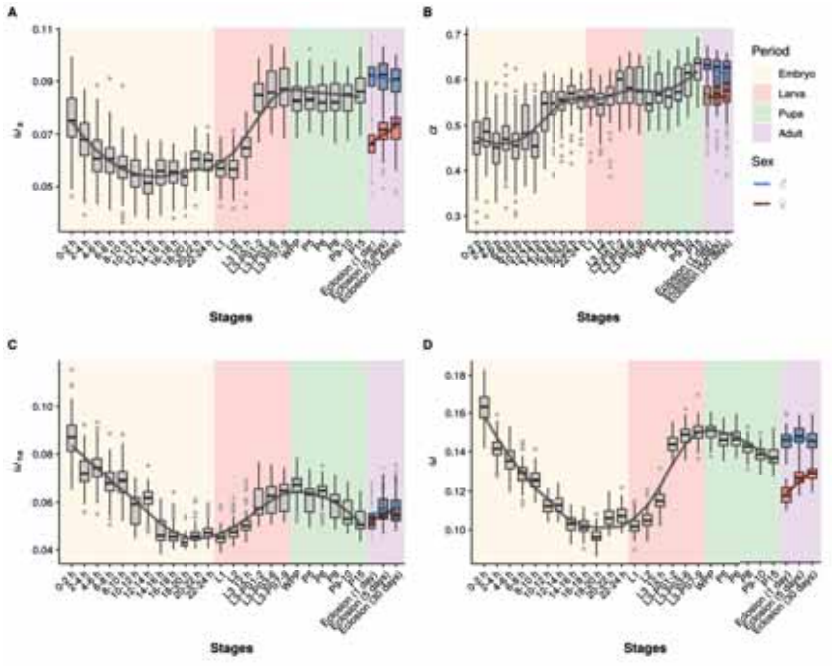


Figure A.8 The selection statistics follow a similar temporal pattern when considering genes expressed with a RPKM $\geqslant$ 2 and using DFE-alpha. In the medium stringent criterion, only genes that have two or more RPKM in that stage are considered. Similar results are found when compared to the pattern of Figure 3.9. **A**. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **B**. α, the proportion of base substitutions fixed by natural selection. **C**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **D**. ω, the rate of non-synonymous substitutions relative to the mutation rate. Colors and lines as in Figure 3.9. The number of analyzed genes is shown in Table B.6.

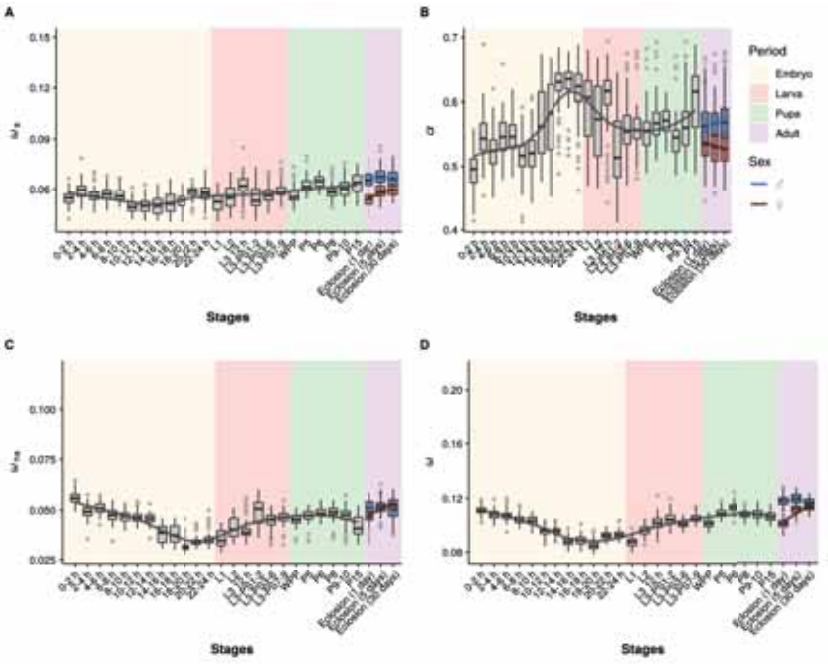


Figure A.9 The selection statistics follow a similar temporal pattern when considering genes expressed with a RPKM $\geqslant$ 10 and using DFE-alpha. In the high stringent criterion, only genes that have ten or more RPKM in that stage are considered. In this case, the stages with maximum and minimum ω_a, ω_{na} and ω are the same that in the previous analyses, but the overall temporal profile is smoother. **A**. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **B**. α, the proportion of base substitutions fixed by natural selection. **C**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **D**. ω, the rate of non-synonymous substitutions relative to the mutation rate. Colors and lines as in Figure 3.9. The number of analyzed genes is shown in Table B.7.

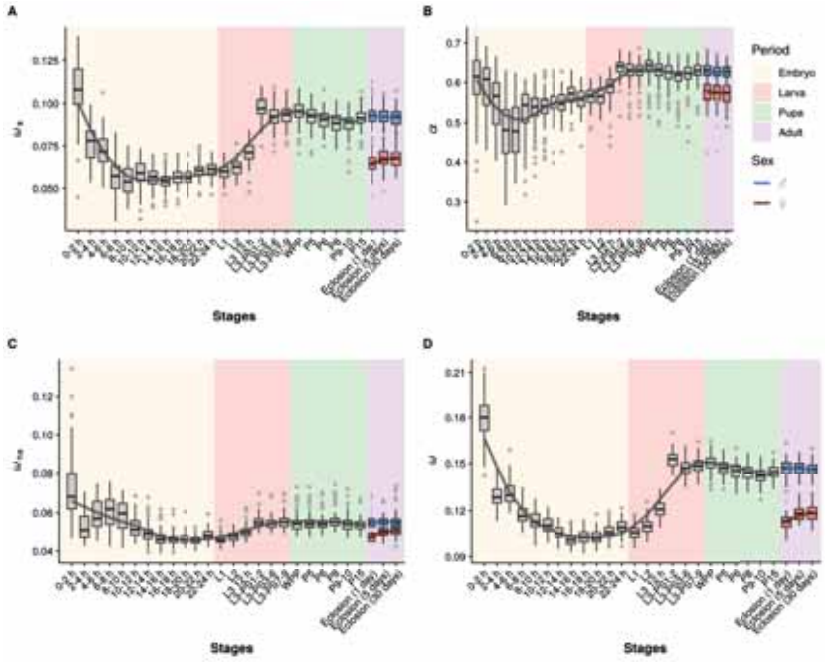


Figure A.10 ω_a, α, ω_{na} and ω over developmental time when genes related with testis and immune are removed using DFE-alpha. **A**. ω_a, the rate of adaptive non-synonymous substitutions relative to the mutation rate. **B**. α, the proportion of base substitutions fixed by natural selection. **C**. ω_{na}, the rate of non-adaptive non-synonymous substitutions relative to the mutation rate. **D**. ω, the rate of non-synonymous substitutions relative to the mutation rate. Colors and lines as in Figure 3.9. The number of analyzed genes is shown in Table B.12.

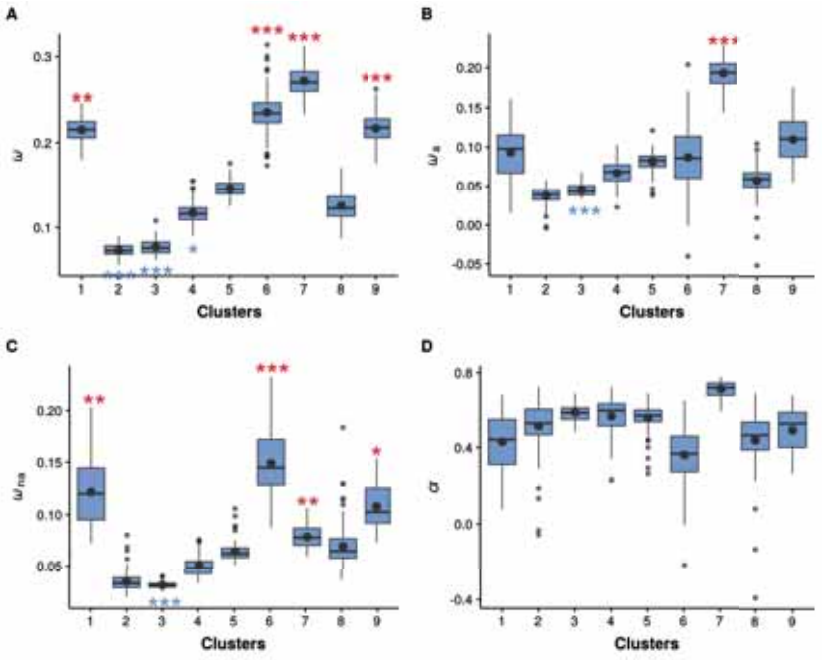


Figure A.11 ω, ω_a, ω_{na} and α for each life cycle cluster. A. ω sampling for each cluster. **B**. ω_a sampling for each cluster. **C**. ω_{na} sampling for each cluster. **D**. α sampling for each cluster. Female data points where discarded and only males were used for the adult stage. Each point in a plot (100 bootstrap replicates per cluster) is calculated for a randomly drawn sample of the set of genes in each cluster with replacement. Number of genes analyzed in Table B.9. Permutation *p*-values are displayed in Table B.19.

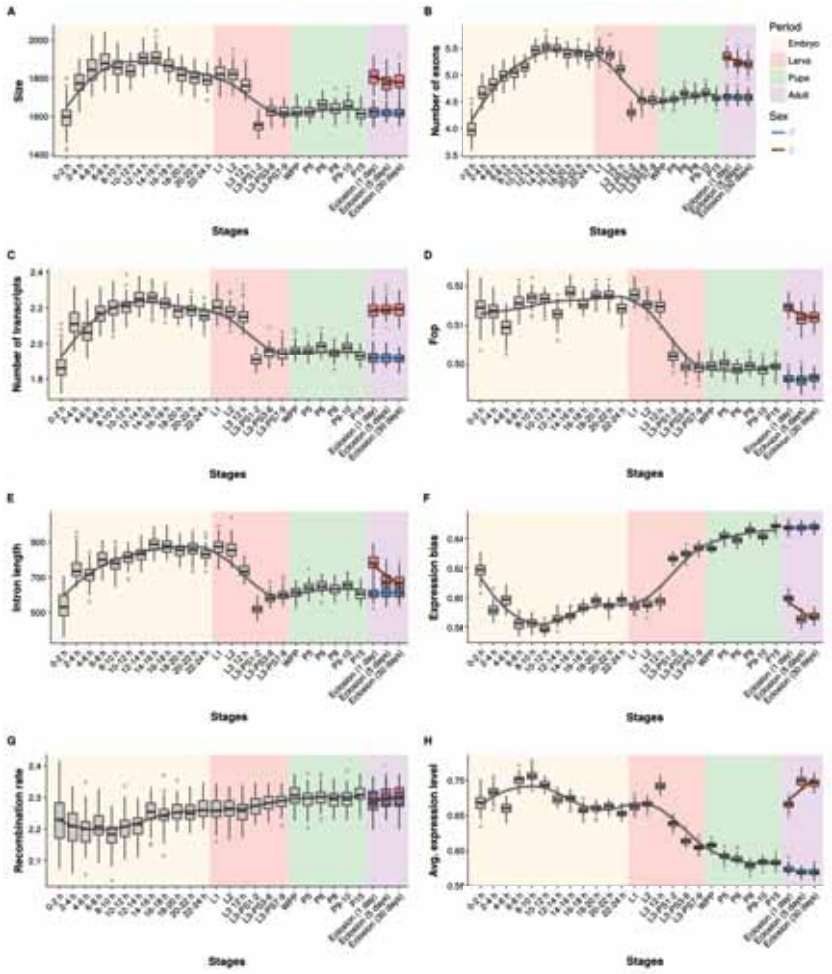


Figure A.12 Six genomic features over developmental stages, using 4-fold data. Lines and stages as in Figure 3.9. **A**. Size is the CDS length of a gene in bp. **B**. Number of exons is the total number of different exons a gene has. **C**. Number of transcripts is the number of different transcripts a gene has. **D**. *Fop* is a measure of codon usage bias, the ratio of optimal codons to synonymous codons. **E**. Intron length is the average distance, in bases, between the exons of a gene. **F**. The expression bias is a measure of how much the expression of a gene is restricted to one or few stages estimated as Equation 2.7. **G**. Recombination rate is estimated in 100 kb non-overlapping windows. **H**. Expression level is the average expression (as the logarithm of the RPKM counts) of a gene in over all stages. Mean sampling distribution obtained by resampling 100 times with replacement the genes from each stage. The number of analyzed genes is shown in Table B.1.

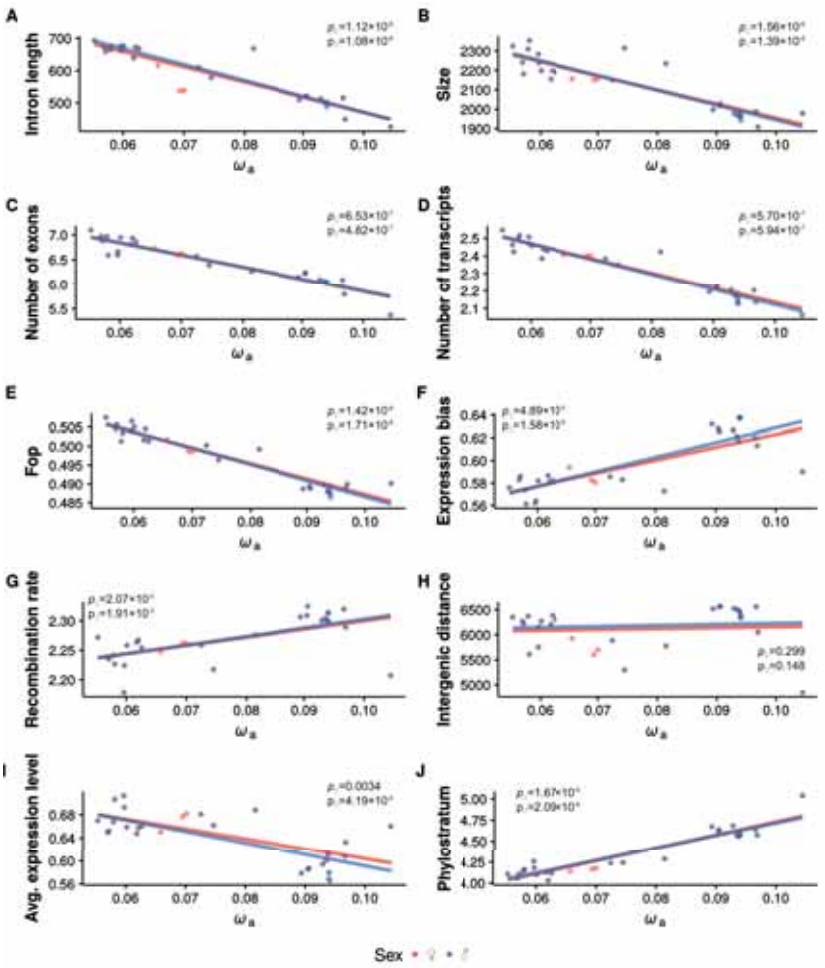


Figure A.13 Correlations between ω_a and the genomic determinants.

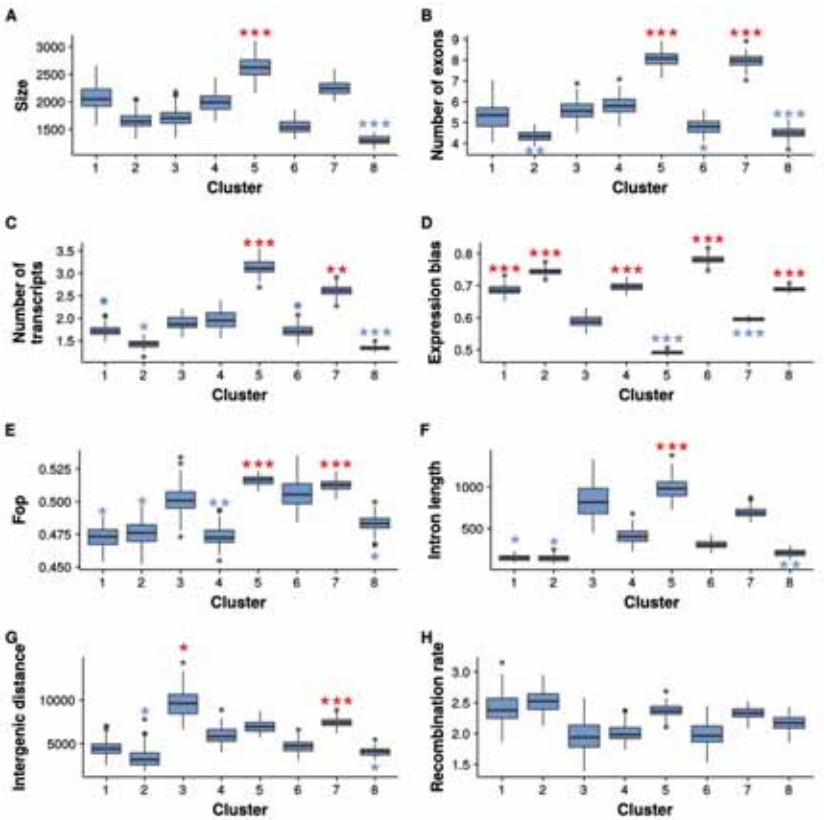


Figure A.14 Genomic features for each cluster of the embryo development. A. Size is the CDS length of a gene in bp. **B**. Number of exons is the total number of different exons a gene has. **C**. Number of transcripts is the number of different transcripts a gene has. **D**. The expression bias is a measure of how much the expression of a gene is restricted to one or few stages estimated as Equation 2.7. **E**. *Fop* is a measure of codon usage bias, the ratio of optimal codons to synonymous codons. **F**. Intron length is the average distance, in bases, between the exons of a gene. **G**. Intergenic distance is the average distance, in bases, between adjacent genes. **H**. Recombination rate is estimated in 100 kb non-overlapping windows. The number of analyzed genes is shown in Table 2.3. Clusters 1, 2 and 8 have significantly lower intron length than other clusters (permutation test, cluster 1: $p = 0.042$; cluster 2: $p = 0.019$; cluster 8: $p = 0.004$, while clusters 5 and 7 show high intron length and high intergenic distance, respectively (permutation test, cluster 5: $p < 0.001$; cluster 7: $p < 0.001$) and rather low ω_a and ω. A similar pattern is found for gene size, number of exons and number of transcripts, that are low for clusters with high ω_a, ω and α (cluster 1, transcripts $p = 0.092$; cluster 2, exons $p = 0.004$, transcripts $p = 0.038$; cluster 8 size $p < 0.001$, exons $p < 0.001$, transcripts $p < 0.001$) and high for clusters 5 and 7 (permutation test, size: cluster 5: $p < 0.001$; transcripts: cluster 5: $p < 0.001$; cluster 7: $p = 0.003$; exons: cluster 5: $p < 0.001$; cluster 7: $p < 0.001$). Codon usage bias, measured as *Fop*, has also a similar pattern with low values in clusters with high ω_a, ω and α (permutation test, cluster 1: $p = 0.019$; cluster 2: $p = 0.027$; cluster 8: $p = 0.019$). Permutation *P*-values are displayed in Table B.21.

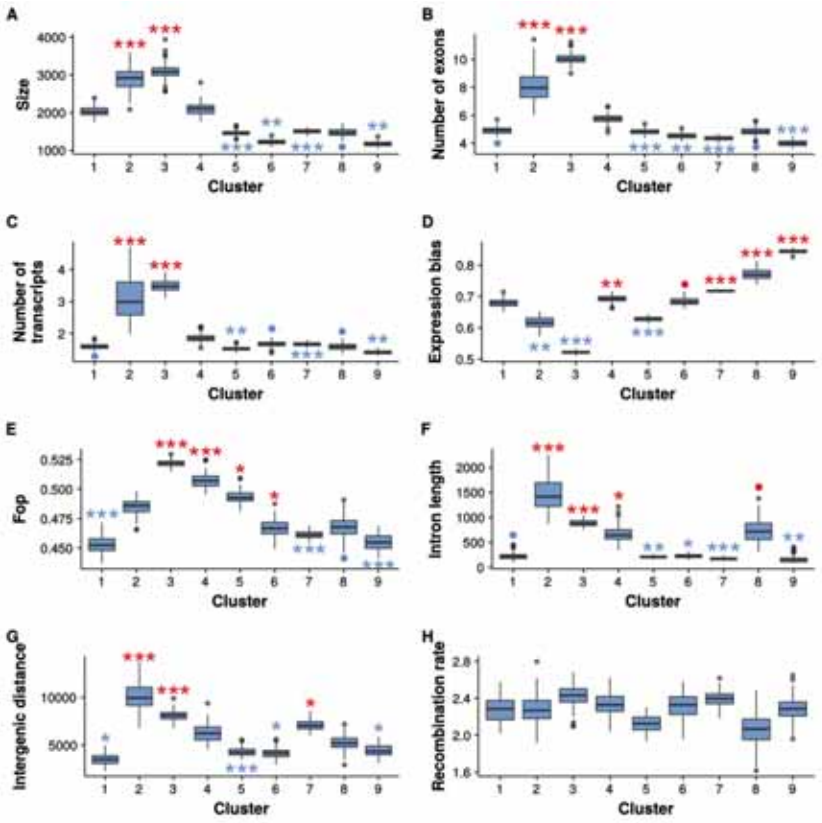


Figure A.15 Genomic features for each cluster of the life cycle. **A**. Size is the CDS length of a gene in bp. **B**. Number of exons is the total number of different exons a gene has. **C**. Number of transcripts is the number of different transcripts a gene has. **D**. The expression bias is a measure of how much the expression of a gene is restricted to one or few stages estimated as Equation 2.7. **E**. *Fop* is a measure of codon usage bias, the ratio of optimal codons to synonymous codons. **F**. Intron length is the average distance, in bases, between the exons of a gene. **G**. Intergenic distance is the average distance, in bases, between adjacent genes. **H**. Recombination rate is estimated in 100 kb non-overlapping windows. The number of analyzed genes is shown in Table B.9. Permutation *p*-values are shown in Table B.22.

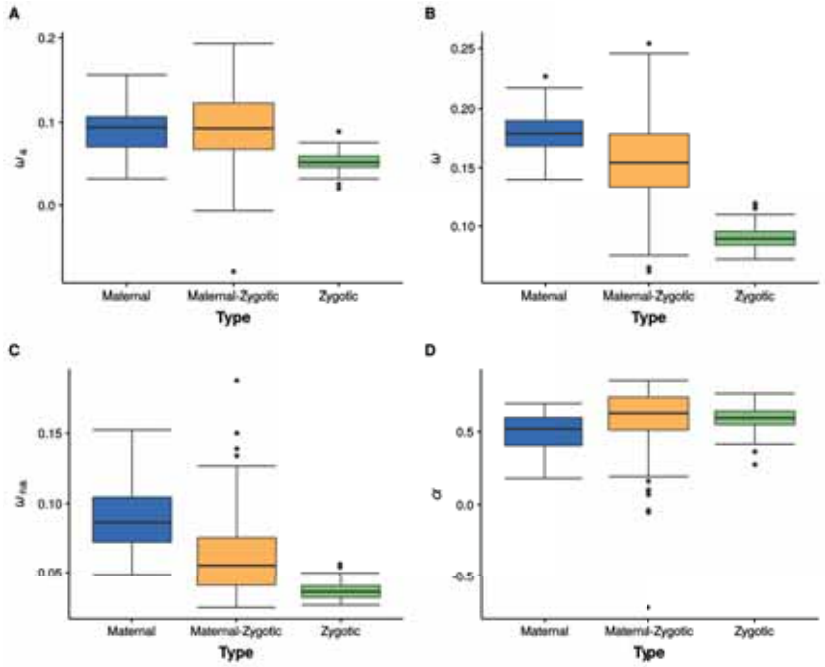


Figure A.16 ω_a, ω, ω_{na} and α for the maternal genes, maternal-zygotic genes and zygotic genes which are in common with the modENCODE dataset. Each boxplot (100 bootstrap replicates per group) is calculated for a randomly drawn sample of the set of genes in each category. Number of genes analyzed: 204 maternal genes, 20 maternal-zygotic and 172 zygotic genes.

B

Supplementary tables

Table B.1 Genes expressed in 30 stages with the low stringent criteria. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Stage	Total genes	4-fold genes	Short-intron genes
0-2 h	1,019	642	333
2-4 h	1,937	1,253	665
4-6 h	2,348	1,466	801
6-8 h	2,729	1,792	989
8-10 h	3,142	2,078	1,154
10-12 h	3,403	2,336	1,320
12-14 h	3,779	2,587	1,488
14-16 h	3,801	2,794	1,572
16-18 h	4,173	2,976	1,679
18-20 h	4,230	3,094	1,735
20-22 h	4,174	3,022	1,686
22-24 h	4,505	3,154	1,771
L1	4,037	2,947	1,648
L2	4,030	2,934	1,645
L3 12 h	3,971	2,889	1,638
L3-PS1-2	5,142	3,737	2,023
L3-PS3-6	5,845	4,067	2,200
L3-PS7-9	6,202	4,165	2,252
WPP	6,002	4,050	2,177
P5	6,491	4,255	2,270
P6	6,497	4,166	2,245
P8	6,837	4,400	2,359
P9-10	6,385	4,234	2,292
P15	6,493	4,480	2,434
♀ Eclosion (1 day)	3,998	3,007	1,750
♀ Eclosion (5 days)	3,424	2,454	1,476
♀ Eclosion (30 days)	3,418	2,468	1,484
♂ Eclosion (1 day)	7,104	4,509	2,484
♂ Eclosion (5 days)	7,164	4,480	2,469
♂ Eclosion (30 days)	7,050	4,483	2,471
Total	9,287	5,323	2,869

Table B.2 Genes expressed in 30 stages with 2-fold change in females. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Stage	Total genes	4-fold genes	Short-intron genes
0-2 h	922	632	326
2-4 h	1,744	1,225	648
4-6 h	2,011	1,411	768
6-8 h	2,413	1,743	958
8-10 h	2,811	2,027	1,124
10-12 h	3,085	2,265	1,270
12-14 h	3,380	2,501	1,423
14-16 h	3,603	2,722	1,522
16-18 h	3,867	2,862	1,601
18-20 h	3,975	2,979	1,659
20-22 h	3,899	2,909	1,616
22-24 h	4,104	3,022	1,687
L1	3,797	2,857	1,594
L2	3,834	2,877	1,611
L3 12 h	3,814	2,853	1,617
L3-PS1-2	4,974	3,686	1,991
L3-PS3-6	5,497	3,978	2,147
L3-PS7-9	5,619	4,044	2,174
WPP	5,467	3,933	2,112
P5	5,716	4,099	2,173
P6	5,541	3,979	2,131
P8	5,821	4,199	2,236
P9-10	5,558	4,066	2,189
P15	5,755	4,263	2,301
♀ Eclosion (1 day)	3,829	2,937	1,706
♀ Eclosion (5 days)	3,193	2,405	1,442
♀ Eclosion (30 days)	3,204	2,412	1,445
Total	6,847	4,836	2,570

Table B.3 Genes expressed in 30 stages with 2-fold change in males. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Stage	Total genes	4-fold genes	Short-intron genes
0-2 h	922	630	325
2-4 h	1,753	1,225	648
4-6 h	2,024	1,413	769
6-8 h	2,430	1,749	963
8-10 h	2,825	2,032	1,127
10-12 h	3,098	2,269	1,272
12-14 h	3,394	2,505	1,426
14-16 h	3,614	2,725	1,525
16-18 h	3,889	2,871	1,607
18-20 h	3,995	2,986	1,664
20-22 h	3,918	2,919	1,624
22-24 h	4,126	3,033	1,695
L1	3,813	2,865	1,599
L2	3,854	2,887	1,618
L3 12 h	3,833	2,865	1,623
L3-PS1-2	5,002	3,701	1,999
L3-PS3-6	5,585	4,001	2,160
L3-PS7-9	5,735	4,070	2,190
WPP	5,565	3,960	2,128
P5	5,879	4,132	2,192
P6	5,723	4,016	2,152
P8	6,002	4,236	2,257
P9-10	5,751	4,103	2,215
P15	6,123	4,374	2,373
♂ Eclosion (1 day)	6,161	4,328	2,373
♂ Eclosion (5 days)	6,077	4,281	2,346
♂ Eclosion (30 days)	6,066	4,278	2,341
Total	7,231	4,925	2,629

Table B.4 Genes expressed in 30 stages with 4-fold change in females. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Stage	Total genes	4-fold genes	Short-intron genes
0-2 h	775	566	292
2-4 h	1,472	1,116	589
4-6 h	1,656	1,274	693
6-8 h	2,033	1,594	879
8-10 h	2,385	1,850	1,028
10-12 h	2,633	2,062	1,160
12-14 h	2,854	2,256	1,286
14-16 h	3,092	2,434	1,372
16-18 h	3,250	2,525	1,428
18-20 h	3,410	2,654	1,487
20-22 h	3,322	2,571	1,437
22-24 h	3,457	2,667	1,496
L1	3,242	2,525	1,417
L2	3,283	2,562	1,438
L3 12 h	3,407	2,651	1,498
L3-PS1-2	4,391	3,377	1,807
L3-PS3-6	4,594	3,529	1,891
L3-PS7-9	4,634	3,567	1,911
WPP	4,510	3,471	1,860
P5	4,681	3,594	1,906
P6	4,487	3,466	1,856
P8	4,741	3,649	1,941
P9-10	4,540	3,528	1,897
P15	4,761	3,710	1,995
♀ Eclosion (1 day)	3,259	2,605	1,516
♀ Eclosion (5 days)	2,748	2,191	1,312
♀ Eclosion (30 days)	2,751	2,189	1,313
Total	5,464	4,144	2,189

Table B.5 Genes expressed in 30 stages with 4-fold change in males. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Stage	Total genes	4-fold genes	Short-intron genes
0-2 h	780	568	293
2-4 h	1,484	1,121	589
4-6 h	1,680	1,287	698
6-8 h	2,056	1,606	884
8-10 h	2,412	1,865	1,033
10-12 h	2,657	2,073	1,162
12-14 h	2,877	2,270	1,290
14-16 h	3,118	2,447	1,376
16-18 h	3,293	2,553	1,440
18-20 h	3,454	2,681	1,498
20-22 h	3,356	2,597	1,450
22-24 h	3,493	2,690	1,508
L1	3,274	2,548	1,427
L2	3,313	2,583	1,448
L3 12 h	3,438	2,670	1,509
L3-PS1-2	4,504	3,448	1,851
L3-PS3-6	4,796	3,645	1,956
L3-PS7-9	4,843	3,682	1,972
WPP	4,711	3,586	1,922
P5	4,902	3,716	1,971
P6	4,714	3,595	1,924
P8	4,970	3,780	2,011
P9-10	4,780	3,664	1,972
P15	5,169	3,917	2,112
♂ Eclosion (1 day)	5,081	3,843	2,093
♂ Eclosion (5 days)	4,995	3,782	2,059
♂ Eclosion (30 days)	5,003	3,792	2,065
Total	5,859	4,334	2,299

Table B.6 Genes expressed in 30 stages with the medium stringent criteria. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Stage	Total genes	4-fold genes	Short-intron genes
0-2 h	1,019	642	333
2-4 h	1,937	1,253	665
4-6 h	2,348	1,466	801
6-8 h	2,729	1,792	989
8-10 h	3,142	2,078	1,154
10-12 h	3,403	2,336	1,320
12-14 h	3,779	2,587	1,488
14-16 h	3,801	2,794	1,572
16-18 h	4,173	2,976	1,679
18-20 h	4,230	3,094	1,735
20-22 h	4,174	3,022	1,686
22-24 h	4,505	3,154	1,771
L1	4,037	2,947	1,648
L2	4,030	2,934	1,645
L3 12 h	3,971	2,889	1,638
L3-PS1-2	5,142	3,737	2,023
L3-PS3-6	5,845	4,067	2,200
L3-PS7-9	6,202	4,165	2,252
WPP	6,002	4,050	2,177
P5	6,491	4,255	2,270
P6	6,497	4,166	2,245
P8	6,837	4,400	2,359
P9-10	6,385	4,234	2,292
P15	6,493	4,480	2,434
♀ Eclosion (1 day)	3,998	3,007	1,750
♀ Eclosion (5 days)	3,424	2,454	1,476
♀ Eclosion (30 days)	3,418	2,468	1,484
♂ Eclosion (1 day)	7,104	4,509	2,484
♂ Eclosion (5 days)	7,164	4,480	2,469
♂ Eclosion (30 days)	7,050	4,483	2,471
Total	9,287	5,323	2,869

Table B.7 Genes expressed in 30 stages with the high stringent criteria. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Stage	Total genes	4-fold genes	Short-intron genes
0-2 h	4,066	3,485	2,449
2-4 h	3,372	2,838	1,937
4-6 h	3,504	2,949	2,010
6-8 h	3,727	3,148	2,145
8-10 h	3,935	3,312	2,231
10-12 h	4,069	3,430	2,274
12-14 h	4,304	3,621	2,456
14-16 h	3,929	3,310	2,155
16-18 h	4,089	3,394	2,206
18-20 h	3,687	3,040	1,905
20-22 h	3,944	3,238	2,073
22-24 h	4,122	3,375	2,182
L1	3,529	2,933	1,847
L2	3,349	2,761	1,719
L3 12 h	3,009	2,452	1,500
L3-PS1-2	2,838	2,326	1,434
L3-PS3-6	4,047	3,344	2,144
L3-PS7-9	4,659	3,835	2,451
WPP	4,450	3,685	2,373
P5	5,203	4,275	2,679
P6	5,670	4,667	2,925
P8	5,691	4,704	2,879
P9-10	4,411	3,641	2,184
P15	4,110	3,382	2,015
♀ Eclosion (1 day)	4,247	3,625	2,367
♀ Eclosion (5 days)	4,700	3,968	2,684
♀ Eclosion (30 days)	4,733	4,006	2,713
♂ Eclosion (1 day)	5,081	4,123	2,484
♂ Eclosion (5 days)	5,456	4,442	2,701
♂ Eclosion (30 days)	5,393	4,439	2,692
Total	10,075	8,042	4,887

Table B.8 Genes expressed in 30 stages with the low stringent criterion. Genes with 0 divergent or polymorphic sites were removed as they cannot be used for estimating α using the MKT method.

Stage	Short-intron genes	Genes with variation
0-2 h	333	242
2-4 h	665	492
4-6 h	801	587
6-8 h	989	739
8-10 h	1,154	872
10-12 h	1,320	1,010
12-14 h	1,488	1,149
14-16 h	1,572	1,217
16-18 h	1,679	1,292
18-20 h	1,735	1,340
20-22 h	1,686	1,302
22-24 h	1,771	1,366
L1	1,648	1,270
L2	1,645	1,266
L312 h	1,638	1,274
L3-PS1-2	2,023	1,566
L3-PS3-6	2,200	1,711
L3-PS7-9	2,252	1,751
WPP	2,177	1,694
P5	2,270	1,751
P6	2,245	1,746
P8	2,359	1,823
P9-10	2,292	1,773
P15	2,434	1,894
♀ Eclosion (1 day)	1,750	1,344
♀ Eclosion (5 days)	1,476	1,137
♀ Eclosion (30 days)	1,484	1,146
♂ Eclosion (1 day)	2,484	1,933
♂ Eclosion (5 days)	2,469	1,913
♂ Eclosion (30 days)	2,471	1,922
Total	2,869	2,232

Table B.9 Genes expressed in 9 clusters of the life cycle. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Cluster	Total genes	Membership $\geqslant 0.8$	4-fold genes	Short-intron genes
1	691	558	276	141
2	707	607	253	116
3	1,128	957	779	472
4	659	521	359	195
5	1,020	890	635	342
6	656	559	325	182
7	2,611	2,496	1,478	722
8	655	491	225	119
9	1,186	1,088	382	232
Total	9,241	8,167	4,712	2,521

Table B.10 Number of genes expressed in each anatomical structures by stage. *4-fold genes* represent genes that can be analyzed with the 4-fold gene dataset and *Short-intron genes* those genes that can be analyzed with the short-intron dataset.

Stage	Anatomical structure	Total genes	4-fold genes	Short-intron genes
1	Germ line#	81	73	47
1	Maternal	4,076	3,593	2,316
1	Ubiquitous	468	408	257
2	Amnioserosa#	135	122	60
2	Ectoderm/Epidermis	291	250	114
2	Endoderm/Midgut#	4	4	1
2	Foregut#	84	66	34
2	Germ line^	168	153	99
2	Hindgut/Malpighian tubules#	69	58	30
2	Maternal	2,335	2,048	1,287
2	Mesoderm/Muscle#	126	109	52
2	Optical lobe#	52	44	22
2	PNS#	1	1	1
2	Procephalic ectoderm/ CNS^	232	189	85
2	Segmental/GAP#	129	114	58
2	Ubiquitous	2,123	1,860	1,180
3	Amnioserosa/Yolk#	117	109	57
3	Ectoderm/Epidermis	373	311	150
3	Endoderm/Midgut	527	459	275
3	Foregut#	102	83	33
3	Germ line^	193	175	109
3	Hindgut/Malpighian tubules^	200	169	93
3	Mesoderm/Muscle	592	512	301
3	Optical lobe#	49	38	20
3	PNS#	5	4	3

Continued on next page

Table B.10 – *Continued from previous page*

Stage	Anatomical structure	Total genes	4-fold genes	Short-intron genes
3	Procephalic ectoderm/ CNS	441	382	212
3	Segmental/GAP#	50	44	16
3	Ubiquitous	2,377	2,081	1,371
4	Amnioserosa/Yolk^	247	220	119
4	Ectoderm/Epidermis	561	476	251
4	Endoderm/Midgut	701	609	382
4	Foregut#	173	142	69
4	Garland/Plasmatocytes/ Ring gland#	34	30	11
4	Germ line#	158	142	87
4	Hindgut/Malpighian tubules^	274	238	137
4	Mesoderm/Muscle	748	649	396
4	Optical lobe#	81	68	31
4	PNS#	19	11	4
4	Procephalic ectoderm/ CNS	470	405	241
4	Salivary gland#	6	4	4
4	Segmental/GAP#	32	29	17
4	SNS#	6	6	1
4	Tracheal system#	8	4	2
4	Ubiquitous	2,283	1,992	1,315
5	Amnioserosa/Yolk	346	304	176
5	Ectoderm/Epidermis	477	401	208
5	Endoderm/Midgut	1,176	1,030	645
5	Foregut	408	355	190
5	Garland/Plasmatocytes/ Ring gland^	257	233	136
5	Germ line#	131	117	74

Continued on next page

Table B.10 – *Continued from previous page*

Stage	Anatomical structure	Total genes	4-fold genes	Short-intron genes
5	Head mesoderm/Circ. Syst./FB	479	422	247
5	Hindgut/Malpighian tubules	592	525	317
5	Mesoderm/Muscle	834	742	445
5	Optical lobe#	54	43	15
5	PNS#	103	79	38
5	Procephalic ectoderm/ CNS	934	803	495
5	Salivary gland#	112	98	68
5	Segmental/GAP#	14	11	5
5	Tracheal system^	306	256	135
5	Ubiquitous	2,034	1,777	1,174
6	Amnioserosa/Yolk	301	262	154
6	Ectoderm/Epidermis	1,119	946	527
6	Endoderm/Midgut	1,596	1,391	861
6	Foregut	849	720	395
6	Garland/Plasmat./ Ring gland	406	358	212
6	Germ line	341	297	209
6	Head mesoderm/Circ. Syst./FB	382	344	192
6	Hindgut/Malpighian tubules	939	826	495
6	Mesoderm/Muscle	903	800	479
6	Optical lobe#	61	48	19
6	PNS^	316	267	145
6	Procephalic ectoderm/ CNS	1,486	1,276	786
6	Salivary gland^	269	238	136
6	Segmental/GAP#	6	4	0

Continued on next page

Table B.10 – *Continued from previous page*

Stage	Anatomical structure	Total genes	4-fold genes	Short-intron genes
6	SNS[#]	52	45	19
6	Tracheal system	432	369	190
6	Ubiquitous	1,599	1,400	918
Total		5,671	4,945	3,028

Anatomical structures not analyzed in posterior analyses (not enough genes to be analyzed, the minimum is 150 genes)

Table B.11 Genomic features analyzed.

Feature	Value (ranks)	Number of genes	Reference
Gene size	[9,617]	2,223	Gramates et al., 2017
	(617,1,020]	2,220	
	(1,020,1,520]	2,219	
	(1,520,2,370]	2,220	
	(2,370,55,400]	2,221	
Number of exons	[1,2]	4,008	Gramates et al., 2017
	(2,3]	1,742	
	(3,4]	1,395	
	(4,7]	2,140	
	(7,114]	1,818	
Number of transcripts	[1]	5,278	Gramates et al., 2017
	[2]	2,775	
	[3]	1,026	
	[4,75]	1,574	
Intron length	[0,54.9]	2,221	Gramates et al., 2017
	(54.9,65]	2,337	
	(65,119]	2,111	
	(119,426]	2,213	
	(426,84,000]	2,221	
Intergenic distance	[0,540]	2,225	Gramates et al., 2017
	(540,1,100]	2,217	
	(1,100,2,570]	2,220	
	(2,570,7,390]	2,220	
	(7,390,184,000]	2,221	
Expression bias	[0.0556,0.249]	2,203	Gelbart and Emmert, 2013
	(0.249,0.354]	2,201	
	(0.354,0.531]	2,202	
	(0.531,0.729]	2,202	
	(0.729,1]	2,202	
Avg. expression	[0,0.408]	2,221	Gelbart and Emmert, 2013

Continued on next page

Table B.11 – *Continued from previous page*

Feature	Value (ranks)	Number of genes	Reference
level	(0.408,0.732]	2,220	
	(0.732,1.06]	2,220	
	(1.06,1.37]	2,220	
	(1.37,3.67]	2,221	
Fop	[0.176,0.441]	2,221	Peden, 1999
	(0.441,0.494]	2,250	
	(0.494,0.538]	2,231	
	(0.538,0.591]	2,187	
	(0.591,0.88]	2,214	
Recombination	[0,0.692]	2,262	Comeron et al., 2,012
rates	(0.692,1.52]	2,243	
	(1.52,2.39]	2,184	
	(2.39,3.59]	2,220	
	(3.59,14.8]	2,194	
Phylogenetic	[1]	3,209	Drost, 2014
age	[2]	2,442	
	[3]	342	
	[4,11]	1,807	
	[12,13]	1,952	

Table B.12 Testis and immune related genes GO terms.

GO code	Definition
GO:0002218	Activation of innate immune response
GO:0002227	Innate immune response in mucosa
GO:0002253	Activation of immune response
GO:0002385	Mucosal immune response
GO:0002433	Immune response-regulating cell surface receptor signaling pathway involved in phagocytosis
GO:0002758	Innate immune response-activating signal transduction
GO:0002775	Antimicrobial peptide production
GO:0002784	Regulation of antimicrobial peptide production
GO:0002805	Regulation of antimicrobial peptide biosynthetic process
GO:0002920	Regulation of humoral immune response
GO:0002921	Negative regulation of humoral immune response
GO:0004766	Spermidine synthase activity
GO:0006597	Spermine biosynthetic process
GO:0006909	Phagocytosis
GO:0006952	Defense response
GO:0006955	Immune response
GO:0006959	Humoral immune response
GO:0006963	Positive regulation of antibacterial peptide biosynthetic process
GO:0006965	Positive regulation of biosynthetic process of antibacterial peptides active against Gram-positive bacteria
GO:0006967	Positive regulation of antifungal peptide biosynthetic process
GO:0007140	Male meiosis
GO:0007283	Spermatogenesis
GO:0007284	Spermatogonial cell division
GO:0007285	Primary spermatocyte growth
GO:0007286	Spermatid development
GO:0007288	Sperm axoneme assembly
GO:0007290	Spermatid nucleus elongation
GO:0007291	Sperm individualization
GO:0007321	Sperm displacement
GO:0007485	Imaginal disc-derived male genitalia development

Continued on next page

Table B.12 – *Continued from previous page*

GO code	Definition
GO:0008584	Male gonad development
GO:0009617	Response to bacterium
GO:0016045	detection of bacterium
GO:0016768	Spermine synthase activity
GO:0019028	Viral capsid
GO:0019730	Antimicrobial humoral response
GO:0019731	Antibacterial humoral response
GO:0030317	Sperm motility
GO:0030382	Sperm mitochondrion organization
GO:0030539	Male genitalia development
GO:0035006	Melanization defense response
GO:0035007	Regulation of melanization defense response
GO:0035009	Negative regulation of melanization defense response
GO:0035041	Sperm chromatin decondensation
GO:0035044	Sperm aster formation
GO:0035260	Internal genitalia morphogenesis
GO:0035323	Male germline ring canal
GO:0036126	Sperm flagellum
GO:0042742	Defense response to bacterium
GO:0045071	Negative regulation of viral genome replication
GO:0045087	Innate immune response
GO:0045088	Regulation of innate immune response
GO:0045089	Positive regulation of innate immune response
GO:0045824	Negative regulation of innate immune response
GO:0046692	Sperm competition
GO:0046693	Sperm storage
GO:0048133	Male germ-line stem cell asymmetric division
GO:0048137	Spermatocyte division
GO:0048515	Spermatid differentiation
GO:0048803	Imaginal disc-derived male genitalia morphogenesis
GO:0050688	Regulation of defense response to virus
GO:0050776	Regulation of immune response

Continued on next page

Table B.12 – *Continued from previous page*

GO code	Definition
GO:0050777	Negative regulation of immune response
GO:0050778	Positive regulation of immune response
GO:0050829	Defense response to Gram-negative bacterium
GO:0050830	Defense response to Gram-positive bacterium
GO:0050832	Defense response to fungus
GO:0050983	Deoxyhypusine biosynthetic process from spermidine
GO:0051533	Positive regulation of NFAT protein import into nucleus
GO:0051607	Defense response to virus
GO:0070725	Yb body
GO:0070864	Sperm individualization complex
GO:0090382	Phagosome maturation
GO:2000019	Negative regulation of male gonad development
GO:2000020	Positive regulation of male gonad development
GO:2000018	Regulation of male gonad development
GO:1990111	Spermatoproteasome complex
GO:0033327	Leydig cell differentiation
GO:0008295	Spermidine biosynthetic process

Table B.13 α and ±SD of the genes in each bin.

Bin	Mean (±SD)	Total	%
1	0.800 (±0.275)	2/1,000	0.2%
2	0.872 (±0.201)	19/3,500	0.54%
5	0.742 (±0.380)	60/2,000	3%
10	0.616 (±0.449)	152/1,000	15.2%
25	0.515 (±0.256)	616/1,000	61.6%
50	0.536 (±0.174)	860/1,000	86%
75	0.546 (±0.138)	930/1,000	93%
100	0.553 (±0.115)	975/1,000	97.5%
250	0.568 (±0.065)	1,000/1,000	100%
500	0.572 (±0.046)	1,000/1,000	100%
750	0.574 (±0.039)	1,000/1,000	100%
1,000	0.575 (±0.034)	1,000/1,000	100%

The *x* cutoff interval is [0,0.9]. A DAF of 20 categories was used.

Table B.14 α and ± estimated in the recombination bins.

Genes	Category	Quantiles	Standard MKT	FWW 5%	FWW 10%	eMKT 5%	eMKT 10%	iMKT
2,776	1	[0,0.653]	-0.258 (±0.041)	0.226 (±0.037)	0.314 (±0.036)	-0.016 (±0.039)	-0.032 (±0.039)	0.435 (±0.037)
2,819	2	(0.653,1.52]	0.122 (±0.032)	0.493 (±0.019)	0.544 (±0.017)	0.334 (±0.024)	0.318 (±0.025)	0.597 (±0.016)
2,743	3	(1.52,2.39]	0.246 (±0.031)	0.559 (±0.02)	0.597 (±0.019)	0.425 (±0.024)	0.407 (±0.025)	0.63 (±0.019)
2,667	4	(2.39,3.58]	0.284 (±0.027)	0.587 (±0.019)	0.623 (±0.018)	0.457 (±0.022)	0.441 (±0.022)	0.66 (±0.017)
2,740	5	(3.58,14.8]	0.37 (±0.026)	0.641 (±0.016)	0.676 (±0.015)	0.53 (±0.02)	0.518 (±0.02)	0.698 (±0.014)

Table B.15 Mean absolute errors between true α values and the estimates from the five MKT approaches.

Simulation	Δstandard	ΔFWW 15%	ΔeMKT 15%	ΔiMKT	ρ_{exp}
Baseline	0.147	0.119	0.124	0.038	0.9
L=10^6	0.148	0.126	0.13	0.075	0.58
L=10^8	0.15	0.122	0.127	0.028	1
T=2×10^4	0.193	0.165	0.17	0.125	0.56
T=2×10^6	0.147	0.119	0.125	0.023	1
μ=10^{-8}	0.13	0.103	0.108	0.023	1
μ=10^{-10}	0.168	0.142	0.146	0.096	0.6
s_d=0.002	0.159	0.133	0.138	0.045	0.94
s_d=0.200	0.127	0.102	0.107	0.044	0.88
r_b=0.0,001	0.202	0.165	0.172	0.048	0.94
r_b=0.0,010	0.11	0.09	0.094	0.035	0.92
s_b=0.02	0.206	0.167	0.174	0.051	0.92
s_b=0.20	0.113	0.091	0.095	0.029	0.88

The increments are estimated as $\Delta_{method}=|\alpha_{method}-\alpha_{true}|$ in each run, averaged over the 50 replicates. ρ_{exp} specifies the proportion of simulations in which the iMKT (exponential) fit was performed.

Table B.16 *P*-value of the permutation test for the periods (using genes expressed in whole development with the low stringent criteria as null distribution). *P*-values corrected with the FDR method.

Period	α	ω_a	ω	ω_{na}
Embryo 0-2h	0.812	0.266	0.032	0.371
Embryo 2-6h	0.812	0.266	0.032	0.250
Embryo 6-24h	0.464	<0.001	<0.001	<0.001
Larva 1-3	0.605	<0.001	<0.001	<0.001
Larva 4-6	0.718	0.782	0.507	0.25
Pupa	0.812	0.605	0.055	0.25
Females	0.812	<0.001	<0.001	<0.001
Males	0.605	0.697	0.272	0.100

Table B.17 *P*-value of the permutation test for the periods (using the whole dataset as null distribution). *P*-values corrected with the FDR method.

Period	α	ω_a	ω	ω_{na}
Embryo 0-2h	0.346	0.002	<0.001	0.637
Embryo 2-6h	0.346	0.110	0.024	0.719
Embryo 6-24h	0.448	0.895	0.221	0.472
Larva 1-3	0.346	0.604	0.840	0.472
Larva 4-6	0.024	<0.001	<0.001	0.719
Pupa	0.056	<0.001	<0.001	0.472
Females	0.346	0.895	0.157	0.472
Males	0.024	<0.001	<0.001	0.713

Table B.18 *P*-value of the permutation test for the clusters. *P*-values corrected with the FDR method.

Cluster	α	ω_a	ω	ω_{na}
1	0.525	0.008	<0.001	0.469
2	0.955	0.059	<0.001	0.120
3	0.376	0.195	0.728	0.230
4	0.525	0.496	0.728	0.256
5	0.376	0.012	<0.001	0.633
6	0.525	0.288	0.265	0.818
7	0.955	0.086	<0.001	0.230
8	0.955	0.086	<0.001	0.230

Table B.19 *P*-value of the permutation test for the clusters in the life cycle. *P*-values corrected with the FDR method.

Cluster	α	ω_a	ω	ω_{na}
1	0.282	0.823	0.008	0.009
2	0.591	0.120	<0.001	0.227
3	0.676	<0.001	<0.001	<0.001
4	0.964	0.425	0.045	0.365
5	0.591	0.526	0.562	0.615
6	0.243	0.622	<0.001	<0.001
7	0.257	<0.001	<0.001	0.016
8	0.470	0.324	0.264	0.621
9	0.658	0.324	<0.001	0.038

Table B.20 Spearman's correlations between adaptation (ω_a) and genomic features.

Genomic variable	Relation with ω_a	Correlation (r^2) in females (p)	Correlation (r^2) in males (p)
Intron length	Negative	0.852 (7.30×10^{-7})	0.772 (1.34×10^{-6})
Gene size	Negative	0.655 (1.99×10^{-6})	0.604 (4.32×10^{-6})
Number of exons	Negative	0.885 (4.90×10^{-7})	0.840 (8.25×10^{-7})
Number of transcripts	Negative	0.725 (1.58×10^{-5})	0.664 (1.88×10^{-6})
Fop	Negative	0.716 (1.61×10^{-6})	0.658 (1.95×10^{-6})
Expression bias	Positive	0.447 (1.95×10^{-4})	0.509 (4.68×10^{-5})
Recombination	Positive	0.233 (1.17×10^{-2})	0.411 (4.26×10^{-4})
Intergenic distance	N.S.	0.050 (0.260)	0.089 (0.131)
Expression level	Negative	0.270 (0.006)	0.386 (7.02×10^{-4})
Phylogenetic age	Positive	0.708 (1.64×10^{-6})	0.699 (1.68×10^{-6})

Spearman's correlations performed between each stage's average ω_a and the average of each genomic feature in each stage. Females and males are separated because their gene expression is measured separately in the last three stages in the modENCODE.

Table B.21 *P*-value of the permutation test for the genomic features of the embryo development clusters. *P*-values corrected with the FDR method.

Category	Gene size	Intron length	Number of exons	Recombination rate	*Fop*	Number of transcripts	Intergenic distance	Expression bias
1	0.994	0.042	0.101	0.678	0.019	0.092	0.271	<0.001
2	0.162	0.019	0.004	0.487	0.027	0.038	0.030	<0.001
3	0.311	0.201	0.183	0.392	0.717	0.224	0.019	0.100
4	0.771	0.201	0.183	0.376	0.008	0.171	0.936	<0.001
5	<0.001	<0.001	<0.001	0.487	<0.001	<0.001	0.146	<0.001
6	0.123	0.114	0.016	0.376	0.374	0.083	0.311	<0.001
7	0.248	0.139	<0.001	0.557	<0.001	0.003	<0.001	<0.001
8	<0.001	0.004	<0.001	0.487	0.019	<0.001	0.016	<0.001

Table B.22 *P*-value of the permutation test for the genomic features of the life cycle clusters. *P*-values corrected with the FDR method.

Category	Gene size	Intron length	Number of exons	Recombination rate Fop	Number of transcripts	Intergenic distance	Expression bias
1	0.424	0.054	0.071	0.991	<0.001	0.060	0.012	0.256
2	<0.001	<0.001	<0.001	0.991	0.834	<0.001	<0.001	0.005
3	<0.001	<0.001	<0.001	0.313	<0.001	<0.001	<0.001	<0.001
4	0.176	0.030	0.961	0.991	<0.001	0.217	0.947	0.015
5	<0.001	0.009	<0.001	0.313	0.027	0.002	<0.001	<0.001
6	0.002	0.049	0.006	0.991	0.026	0.060	0.012	0.088
7	<0.001	<0.001	<0.001	0.313	<0.001	<0.001	0.011	<0.001
8	0.071	0.051	0.055	0.313	0.090	0.053	0.298	<0.001
9	0.002	0.009	<0.001	0.991	<0.001	0.002	0.021	<0.001

Table B.23 *P*-values of the permutation test for maternal, maternal-zygotic and zygotic genes. *P*-values corrected with the FDR method.

Class	α	ω_a	ω	ω_{na}
Maternal	0.075	0.920	0.024	0.003
Maternal-Zygotic	0.603	0.920	0.156	0.246
Zygotic	0.249	0.920	0.035	0.036

Table B.24 Recombination rate average levels in each germ layer and statistical analysis.

Layer	Recombination average (±SD)	Homogeneity of variances (Fligner-Killeen test *p*-value)	*p*-value corrected FDR	ANOVA *p*-value	*p*-value corrected FDR
Ectoderm	2.200 (±1.922)	0.043	n.s.	0.091	n.s.
Endoderm	2.326 (±2.020)	0.7	n.s.	0.737	n.s.
Mesoderm	2.194 (±1.877)	0.08	n.s.	0.422	n.s.

The comparisons in the ANOVA/Fligner-Killen test are done using the anatomical structure dataset as a reference (5,969 genes, from which 5,165 genes are analyzed with the gene dataset).

Table B.25 Mutation rate (K_{4f}) average levels in each germ layer and statistical analysis.

Layer	Mutation rate average (±SD)	Homogeneity of variances (Fligner-Killeen test *p*-value)	*p*-value corrected FDR	ANOVA *p*-value	FDR *p*-value
Ectoderm	0.167 (±0.054)	0.124	n.s.	$4.86\times^{-3}$	$7.29\times^{-3}$
Endoderm	0.188 (±0.054)	0.641	n.s.	$1.00\times^{-6}$	$3.01\times^{-6}$
Mesoderm	0.170 (±0.067)	0.708	n.s.	0.698	n.s.

The comparisons in the ANOVA/Fligner-Killen test are done using the anatomical structure dataset as a reference (5,969 genes, from which 5,165 genes are analyzed with the gene dataset).

Table B.26 Gene density average in each germ layer and statistical analysis.

Layer	Density average (±SD)	Homogeneity of variances (Fligner-Killeen test *p*-value)	FDR *p*-value	ANOVA *p*-value	FDR *p*-value
Ectoderm	28,236.99 (±15,474.87)	0.183	n.s.	0.037	n.s.
Endoderm	29,389.59 (±13,994.62)	0.153	n.s.	0.708	n.s.
Mesoderm	28,682.9 (±13,526.86)	0.111	n.s.	0.674	n.s.

The comparisons in the ANOVA/Fligner-Killen test are done using the anatomical structure dataset as a reference (5,969 genes, from which 5,165 genes are analyzed with the gene dataset).

Table B.27 Permutation test *p*-value for anatomical structures analyzed with short-intron sites.

Anatomical structure	ω	ω_a	ω_{na}	α
Amnioserosa/Yolk	n.s.	n.s.	n.s.*	n.s.
Ectoderm/Epidermis	<0.001	0.066	n.s.*	n.s.
Endoderm/Midgut	0.004	n.s.*	n.s.	n.s.
Foregut	0.004	0.006	n.s.	0.020
Garland/Plasmat./ Ring gland	n.s.*	0.020	n.s.	0.042
Germ line	n.s.*	n.s.*	n.s.	n.s.
Head mesoderm/Circ. Syst./FB	n.s.	n.s.	n.s.	n.s.
Hindgut/Malpighian tubules	0.012	n.s.*	n.s.	n.s.
Maternal	n.s.	n.s.	n.s.	n.s.
Mesoderm/Muscle	0.004	n.s.*	n.s.	n.s.
PNS	<0.001	n.s.*	n.s.	n.s.
Procephalic ectoderm/ CNS	0.004	n.s.	0.020	0.072
Salivary gland	n.s.*	0.016	n.s.*	0.030
Tracheal system	n.s.	n.s.	n.s.	n.s.
Ubiquitous	n.s.	n.s.	n.s.	n.s.

*Marginally significant when *p*-value is 1 tailed. Only analyzed anatomical structures with more than 150 genes.

Table B.28 Recombination rate average levels in anatomical structure and statistical analysis.

Anatomical term	Recombination average (±SD)	Homogeneity of variances (Fligner-Killeen test *p*-value)	*p*-value corrected FDR	ANOVA *p*-value	*p*-value corrected FDR
Amnioserosa/Yolk	2.210 (±1.855)	0.438	n.s.	0.397	n.s.
Ectoderm/Epidermis	2.265 (±1.891)	0.491	n.s.	0.642	n.s.
Endoderm/Midgut	2.315 (±1.961)	0.488	n.s.	0.488	n.s.
Foregut	2.269 (±1.929)	0.522	n.s.	0.749	n.s.
Garland/Plasmat./ Ring gland	2.261 (±1.900)	0.253	n.s.	0.778	n.s.
Germ line	2.264 (±1.987)	0.772	n.s.	0.798	n.s.
Head mesoderm/Circ. Syst./FB	2.184 (±1.932)	0.228	n.s.	0.155	n.s.
Hindgut/Malpighian tubules	2.278 (±1.875)	0.226	n.s.	0.862	n.s.
Maternal	2.311 (±1.967)	0.176	n.s.	0.149	n.s.
Mesoderm/Muscle	2.285 (±1.955)	0.945	n.s.	0.959	n.s.
Optical lobe	2.467 (±2.241)	0.317	n.s.	0.297	n.s.
PNS	2.199 (±1.929)	0.602	n.s.	0.423	n.s.
Procephalic ectoderm/ CNS	2.290 (±1.975)	0.666	n.s.	0.945	n.s.
Salivary gland	2.137 (±1.726)	0.053	n.s.	0.183	n.s.
Segmental/GAP	2.161 (±2.030)	0.786	n.s.	0.455	n.s.
SNS	1.899 (±1.740)	0.518	n.s.	0.173	n.s.
Tracheal system	2.291 (±1.947)	0.848	n.s.	0.969	n.s.
Ubiquitous	2.336 (±1.996)	0.032	n.s.	0.071	n.s.

The comparisons in the ANOVA/Fligner-Killen test are done using the anatomical structure dataset as a reference (5,969 genes, from which 5,165 genes are analyzed with the gene dataset).

Table B.29 Mutation rate (K_{4f}) average levels in anatomical structure and statistical analysis.

Anatomical term	Mutation average (±SD)	Homogeneity of variances (Fligner-Killeen test *p*-value)	*p*-value corrected FDR	ANOVA *p*-value	FDR *p*-value
Amnioserosa/Yolk	0.167 (±0.054)	0.185	n.s.	0.119	n.s.
Ectoderm/Epidermis	0.159 (±0.055)	0.059	n.s.	1.31×10^{-20}	2.35×10^{-19}
Endoderm/Midgut	0.169 (±0.055)	0.09	n.s.	0.085	n.s.
Foregut	0.159 (±0.055)	0.439	n.s.	8.71×10^{-14}	5.22×10^{-13}
Garland/Plasmat./ Ring gland	0.160 (±0.056)	0.102	n.s.	2.08×10^{-5}	3.74×10^{-5}
Germ line	0.175 (±0.058)	0.985	n.s.	0.155	n.s.
Head mesoderm/Circ. Syst./FB	0.169 (±0.057)	0.805	n.s.	0.261	n.s.
Hindgut/Malpighian tubules	0.163 (±0.053)	0.002	0.0,309	5.65×10^{-7}	1.27×10^{-6}
Maternal	0.174 (±0.056)	0.358	n.s.	1.54×10^{-6}	3.08×10^{-6}
Mesoderm/Muscle	0.163 (±0.058)	0.535	n.s.	6.53×10^{-9}	2.35×10^{-8}
Optical lobe	0.145 (±0.053)	0.594	n.s.	1.32×10^{-7}	3.96×10^{-7}
PNS	0.155 (±0.053)	0.122	n.s.	1.73×10^{-7}	4.46×10^{-7}
Procephalic ectoderm/ CNS	0.161 (±0.058)	0.066	n.s.	1.67×10^{-17}	1.50×10^{-16}
Salivary gland	0.167 (±0.056)	0.342	n.s.	0.225	n.s.
Segmental/Gap	0.152 (±0.059)	0.199	n.s.	5.64×10^{-5}	9.23×10^{-5}
SNS	0.147 (±0.060)	0.386	n.s.	3.61×10^{-3}	5.42×10^{-3}
Tracheal system	0.156 (±0.053)	0.018	n.s.	2.34×10^{-10}	1.06×10^{-9}
Ubiquitous	0.172 (±0.057)	0.375	n.s.	0.42	n.s.

The comparisons in the ANOVA/Fligner-Killen test are done using the anatomical structure dataset as a reference (5,969 genes, from which 5,165 genes are analyzed with the gene dataset).

Table B.30 Density average levels in anatomical structure and statistical analysis.

Anatomical term	Density average (±SD)	Homogeneity of variances (Fligner-Killeen test *p*-value)	*p*-value corrected FDR	ANOVA *p*-value	FDR *p*-value
Amnioserosa/Yolk	28,276.67 (±15,252.51)	0.827	n.s.	0.268	n.s.
Ectoderm/Epidermis	28,198.58 (±15,736.26)	0.017	n.s.	0.018	0.04
Endoderm/Midgut	29,712.36 (±14,777.18)	0.976	n.s.	0.031	n.s.
Foregut	28,928.85 (±15,510.51)	0.094	n.s.	0.779	n.s.
Garland/Plasmat./ Ring gland	28,833.39 (±14,653.44)	0.894	n.s.	0.753	n.s.
Germ line	29,395.97 (±14,646.84)	0.88	n.s.	0.628	n.s.
Head mesoderm/Circ. Syst./FB	28,783.69 (±14,743.77)	0.822	n.s.	0.624	n.s.
Hindgut/Malpighian tubules	30,064.67 (±15,434.71)	0.039	n.s.	0.016	0.042
Maternal	29,643.09 (±14,847.18)	0.307	n.s.	4.04×10^{-6}	7.28×10^{-5}
Mesoderm/Muscle	29,198.36 (±14,949.43)	0.658	n.s.	0.697	n.s.
Optical lobe	24,602.59 (±14,572.78)	0.895	n.s.	7.44×10^{-4}	4.46×10^{-3}
PNS	28,647.39 (±16,165.25)	0.061	n.s.	0.629	n.s.
Procephalic ectoderm/ CNS	28,483.50 (±15,490.24)	0.019	n.s.	0.081	n.s.
Salivary gland	30,769.15 (±15,004.98)	0.737	n.s.	0.047	n.s.
Segmental/Gap	24,976.21 (±16,523.63)	0.618	n.s.	1.69×10^{-3}	7.62×10^{-3}
SNS	23,067.53 (±14,686.63)	0.167	n.s.	6.00×10^{-3}	0.018
Tracheal system	27,136.9 0(±16,268.12)	0.127	n.s.	3.09×10^{-3}	0.011
Ubiquitous	29,906.60 (±15,058.90)	0.639	n.s.	4.37×10^{-5}	3.93×10^{-4}

The comparisons in the ANOVA/Fligner-Killen test are done using the anatomical structure dataset as a reference (5,969 genes, from which 5,165 genes are analyzed with the gene dataset).

Table B.31 *P*-values of the permutation test. 4-fold sites are used as a proxy for the mutation rate.

Stage	Anatomical structure	ω	ω_a	ω_{na}	α
1	Maternal	n.s.	n.s.	0.002	0.010
1	Ubiquitous	n.s.	n.s.	n.s.	n.s.
2	Ectoderm/Epidermis	0.054	n.s.	n.s.	n.s.
2	Germ line	0.012	0.010	n.s.	n.s.
2	Maternal	n.s.	n.s.	n.s.	n.s.
2	Procephalic ectoderm/ CNS	0.001	0.074	n.s.	n.s.
2	Ubiquitous	n.s.	n.s.	n.s.	n.s.
3	Ectoderm/Epidermis	0.002	0.056	n.s.	n.s.
3	Endoderm/Midgut	0.004	n.s.	0.001	n.s.
3	Germ line	0.008	0.002	n.s.	0.004
3	Hindgut/Malpighian tubules	0.046	n.s.	n.s.	n.s.
3	Mesoderm/Muscle	n.s.	n.s.	0.094	n.s.
3	Procephalic ectoderm/ CNS	n.s.	0.094	n.s.	0.096
3	Ubiquitous	0.010	0.076	n.s.	n.s.
4	Amnioserosa/Yolk	n.s.	n.s.	n.s.	n.s.
4	Ectoderm/Epidermis	0.020	n.s.	n.s.	n.s.
4	Endoderm/Midgut	0.014	n.s.	0.001	n.s.
4	Hindgut/Malpighian tubules	0.068	n.s.	0.022	n.s.
4	Mesoderm/Muscle	0.078	n.s.	0.014	n.s.
4	Procephalic ectoderm/ CNS	n.s.	n.s.	n.s.	n.s.
4	Ubiquitous	0.006	n.s.	n.s.	n.s.
5	Amnioserosa/Yolk	n.s.	n.s.	n.s.	n.s.
5	Ectoderm/Epidermis	0.001	0.068	0.034	n.s.
5	Endoderm/Midgut	0.001	n.s.	0.002	n.s.
5	Foregut	n.s.	n.s.	0.002	0.070

Continued on next page

Table B.31 – *Continued from previous page*

Stage	Anatomical structure	ω	ω_a	ω_{na}	α
5	Garland/Plasmat./ Ring gland	n.s.	n.s.	n.s.	n.s.
5	Head mesoderm/Circ. Syst./FB	n.s.	n.s.	0.012	0.094
5	Hindgut/Malpighian tubules	0.030	n.s.	0.012	n.s.
5	Mesoderm/Muscle	n.s.	n.s.	0.092	n.s.
5	Procephalic ectoderm/ CNS	0.014	n.s.	0.090	n.s.
5	Tracheal system	n.s.	n.s.	0.040	0.066
5	Ubiquitous	0.004	0.048	n.s.	n.s.
6	Amnioserosa/Yolk	n.s.	n.s.	n.s.	n.s.
6	Ectoderm/Epidermis	0.001	0.001	0.032	n.s.
6	Endoderm/Midgut	0.001	0.012	n.s.	n.s.
6	Foregut	0.001	n.s.	0.058	n.s.
6	Garland/Plasmat./ Ring gland	0.014	0.016	0.058	0.006
6	Germ line	0.016	0.014	n.s.	n.s.
6	Head mesoderm/Circ. Syst./FB	n.s.	0.008	n.s.	0.014
6	Hindgut/Malpighian tubules	0.002	0.072	0.066	n.s.
6	Mesoderm/Muscle	0.078	n.s.	n.s.	n.s.
6	PNS	0.001	0.016	n.s.	n.s.
6	Procephalic ectoderm/ CNS	0.001	n.s.	0.001	n.s.
6	Salivary gland	0.001	0.001	n.s.	0.012
6	Tracheal system	n.s.	n.s.	n.s.	n.s.
6	Ubiquitous	0.046	n.s.	n.s.	n.s.

Table B.32 Permutation test *p*-values for anatomical structures. Short-intron sites are used as a proxy for the mutation test.

Stage	Anatomical structure	ω	ω_a	ω_{na}	α
1	Maternal	n.s.	n.s.	n.s.	n.s.
1	Ubiquitous	n.s.	n.s.	n.s.	n.s.
2	Maternal	n.s.	n.s.	n.s.	n.s.
2	Ubiquitous	n.s.	n.s.	n.s.	n.s.
3	Ectoderm/Epidermis	0.038	n.s.	n.s.	n.s.
3	Endoderm/Midgut	0.006	n.s.	n.s.	n.s.
3	Mesoderm/Muscle	n.s.	n.s.	0.044	n.s.
3	Procephalic ectoderm/CNS	n.s.	n.s.	n.s.	n.s.
3	Ubiquitous	n.s.	n.s.	n.s.	n.s.
4	Ectoderm/Epidermis	0.074	n.s.	0.056	n.s.
4	Endoderm/Midgut	0.056	n.s.	n.s.	n.s.
4	Mesoderm/Muscle	0.05	n.s.	0.002	0.054
4	Procephalic ectoderm/CNS	n.s.	n.s.	n.s.	n.s.
4	Ubiquitous	n.s.	n.s.	n.s.	n.s.
5	Amnioerosa/Yolk	n.s.	n.s.	n.s.	n.s.
5	Ectoderm/Epidermis	0.001	0.084	n.s.	n.s.
5	Endoderm/Midgut	0.05	n.s.	n.s.	n.s.
5	Foregut	n.s.	n.s.	n.s.	n.s.
5	Garland/Plasmat./Ring gland	n.s.	n.s.	n.s.	n.s.
5	Head mesoderm/Circ. syst./FB	n.s.	n.s.	n.s.	n.s.
5	Hindgut/Malpighian tubules	n.s.	n.s.	n.s.	n.s.
5	Mesoderm/Muscle	0.034	n.s.	n.s.	n.s.
5	Procephalic ectoderm/CNS	0.080	n.s.	n.s.	n.s.
5	Ubiquitous	n.s.	n.s.	n.s.	n.s.
6	Amnioserosa/Yolk	0.016	n.s.	n.s.	n.s.
6	Ectoderm/Epidermis	0.001	0.002	n.s.	n.s.
6	Endoderm/Midgut	0.066	n.s.	n.s.	n.s.
6	Foregut	0.006	0.016	n.s.	0.038
6	Garland/Plasmat./Ring gland	0.068	0.052	n.s.	n.s.
6	Germ line	n.s.	n.s.	n.s.	n.s.
6	Head mesoderm/Circ. Syst./FB	n.s.	n.s.	n.s.	n.s.
6	Hindgut/Malpighian tubules	n.s.	n.s.	n.s.	n.s.
6	Mesoderm/Muscle	0.006	0.040	n.s.	n.s.
6	Procephalic ectoderm/CNS	0.002	n.s.	0.034	n.s.
6	Tracheal system	n.s.	n.s.	n.s.	n.s.
6	Ubiquitous	n.s.	n.s.	n.s.	n.s.

Only analyzed anatomical structures with more than 150 genes.

Colophon

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". The bibliography was processed by Biblatex. `classicthesis` is available for both LaTeX and LyX:

`https://bitbucket.org/amiede/classicthesis/`

Palatino and *Euler* typefaces are used in the main text. Monospaced text is typeset in *Bera Mono*. *Source Sans Pro* acts as both tables and captions typeface.