Estimating complexity and adaptation in the embryo: a statistical developmental biology approach

Irepan Salvador-Martínez

Institute of Biotechnology

and

Division of Genetics
Department of Biosciences
Faculty of Biological and Environmental Sciences

and

Doctoral Programme in Integrative Life Science
University of Helsinki

ACADEMIC DISSERTATION

To be presented in ... text of a long permission notice. Text of a long permission notice. Text of a long permission notice. Text of a long permission notice. Text of a long permission notice. Text of a long permission notice.

Helsinki 2016

Supervisor
Isaac Salazar-Ciudad, University of Helsinki, Finland

Pre-examiners
Pavel Tomancak, Max Planck Institute of Molecular Cell Biology and Genetics in Dresden, Germany
Gregor Bucher, Georg-August-University Göttingen, Germany

Opponent
Johannes Jäger, Konrad Lorenz Institute, Austria

Custos
Name, University, Country

Advisory Committee
Jukka Jernvall, University of Helsinki
Osamu Shimmi, University of Helsinki
Mikael Fortelius, University of Helsinki

ISSN 1238-8645
ISBN 000-00-0000-0 (paperback)
ISBN 000-00-0000-0 (PDF)
Helsinki 2016
Unigrafia

Acknowledgements

Contents

List of publications

Abbreviations

α	Proportion of adaptive nucleotide substitutions
ω	dN/dS ratio
ω_α	Proportion of adaptive non-synonymous substitutions
Dn	Non-synonymous (inter-specific) divergence per site
Ds	Synonymous (inter-specific) divergence per site
N_e	Effective population size
Pn	Synonymous (intra-specific) polymorphism per site
Ps	Synonymous (intra-specific) polymorphism per site
s	Selection coefficient
2D	two-dimensional
3D	three-dimensional
A/P	anterior/posterior
ANOVA	Analysis of Variance
BDGP	Berkeley Drosophila Genome Project
CNS	Central Nervous System
D/V	dorsal/ventral
DFE	Distribution of Fitness Effects
DGRP	The *Drosophila melanogaster* Genetic Reference Panel
DNA	Deoxyribonucleic acid
DNE	Dirichlet Normal Energy
evo-devo	Evolutionary developmental biology
GF	Growth Factor
GIS	Geographic Information Systems
GRN	Gene Regulatory Network
HG	Hourglass model
IQR	Inter Quartile Range
KW	Kruskal-Wallis test

miRNAs Micro RNAs

MKT MacDonald-Kreitman test

PCA Principal Component Analysis

RNA Ribonucleic acid

RTK receptor tyrosine kinase

SEM Standard error of the mean

SFS Site Frequency Spectrum

SIGs Signaling molecules

siRNAs Small interference RNAs

SNP Single nucleotide polymorphism

ST Synexpression territories

Abstract

Embryonic development has amazed scientists for centuries. Many reasons have been suggested for the perceivable increase in complexity in development, during which a single cell into a larva or an adult. At the level of gene expression, it its assumed that genes change from being expressed in large spatial domains of the embryo in early development to spatially restricted domains (e.g., tissues, cells) in late development. For many developmental genes, the spatio-temporal expression dynamics has been thoroughly described. It is not clear however, if the global dynamics are similar, or if there are differences between types of genes or between species.

Adaptive reasons have been also said to be the cause for the increase in complexity. Adaptations could be estimated with molecular evolution methods based on the analysis of genes expressed in different developmental stages or regions in the embryo. These methods estimate adaptive changes at the DNA sequence level assuming that a positive selected site would show less variance than other sites evolving neutrally. Different developmental stages might show distinct levels of positive or stabilizing selection, that could be related to inter-specific divergence patterns proposed by the von Baer's laws or the hourglass model. The former states that the development of two species of a phylogenetic group would be very similar in early stages and increasingly divergent in subsequent stages. In contrast, the latter states that development is less divergent (more conserved) at mid development.

In here, I analysed gene expression information to estimate both complexity and adaptation in the embryo using a statistical approach. To measure complexity, I developed quantitative measures of spatial complexity and used them in publicly available gene expression data (thousands of in situ hybridization experiments) in *Drosophila melanogaster* and *Ciona intestinalis* from the BDGP/FlyExpress and ANISEED databases respectively. To estimate adaptation, I combined diverse *D. melanogaster* gene expression data (modENCODE, in situs from the BDGP/FlyExpress and gene expression data based on a controlled vocabulary of the embryo anatomy) with population genomic data (from the DGRP project). Using the DFE-alpha method (which uses coding-region polymorphism and divergence to estimate the proportion of adaptive changes) , I charted a spatial map on adaptation of the fruit fly embryo's anatomy. Finally, I analysed the pattern of positive selection through the entire life cycle of *D. melanogaster* and how it correlated with specific genomic determinants (e.g., gene structure, codon bias)

Briefly, I found that *Drosophila* and *Ciona* complexity increases non-linearly with the major change in complexity being before and after gastrulation, respectively. In both species, transcription factors and signalling molecules showed an earlier compartmentalization, consistent with their proposed leading role in pattern formation. In *Drosophila*, gonads and head showed high adaptation during embryogenesis, although pupa and adult male stages exhibit the highest levels of adaptive change, and mid and late embryonic stages show high conservation, showing an HG pattern. Furthermore, I propose that the explanation for the lack of conservation in the early stages could be a relaxation of natural selection, and that the hourglass pattern could be explained by gene structure complexity.

Review of the literature

During the last decades the scientific community has witnessed the flourishing of modern developmental biology (although developmental biology can not be considered a young scientific discipline, as its roots come from centuries ago from embryology and anatomy). Since the 1980's crucial discoveries (Gilbert, 1998) have improved our understanding of the developmental process in many model organisms.

Most of the modern developmental biology studies use an "individualistic" approach (Davidson, 2009), e.g., focusing only on the description of some gene's effect on the development of a specific structure or the role of a gene in a specific signalling pathway. This individualistic approach has increased substantially the knowledge in the developmental biology field and has accumulated a great amount of gene expression information in many years of collective efforts of the developmental biology community. The emergence of methods like DNA microarrays extended the determination of the expression of a single gene to a genomic level, allowing new systemic approaches to study gene expression during development. An example of the results obtained by these approaches is the identification of groups of temporal co-expressing genes during development (e.g., Arbeitman et al., 2002; Hooper et al., 2007).

The majority of the systemic approaches on gene expression during development have focused in the temporal analysis of expression, without considering the spatial distribution of the expressing genes in the embryo (there are however some noteworthy studies that have analysed the spatial patterns of gene expression during development, e.g., Gurunathan et al., 2004; Tomancak et al., 2007; Frise et al., 2010; Crombach et al., 2012; Konikoff et al., 2012)

The analysis of the spatial patterns of gene expression is now facilitated by recent high-throughput in situ hybridization approaches (Tomancak et al., 2002; Pollet et al., 2003; Imai et al., 2004; Christiansen et al., 2006; Lécuyer et al., 2007; Tassy et al., 2010), which have not only further increase the amount of spatio-temporal gene expression data during development of some model organisms, but also allow for straightforward comparisons between gene expression patterns with computational methods.

The individualistic approach is also common in studies that aim to detect natural selection. Most studies that directly search for adaptation at the phenotypic level analyse only a single trait or a small number of traits. However, there is no study that has estimated natural selection over the entire body of an organism. As any adaptive change in the phenotype is expected to be partially caused by genetic mutation, an alternative to detect natural selection is the analysis of DNA sequences of genes expressing differentially in different parts of an organism's body. This could be extended to different stages in the life cycle of an organism if there is enough spatio-temporal gene expression information.

Therefore, the availability of gene expression at a genomic level allows to shift the focus of developmental biology from the study of single genes to a systemic approach in which the global statistical properties of development can be investigated.

In the next subsections, I will make an introduction of the study of complexity and adaptation during embryonic development, emphasizing the methods and concepts that have been previously (or could be potentially) used to analyse both.

Before I do this, it might be useful to define what is development. So firstly, I will address this apparently simple question.

What is development?

" It is not enough to see that horse pulling a cart past the window as the good working horse it is today; the picture must also include the minute fertilised egg, the embryo in its mother's womb, and the broken-down old nag it will eventually become."

C. H. Waddington 1957

It seems that there is no unique or straightforward answer to this question. Sometimes, the study of development is implicitly considered to be the same as the the study of embryology (Horder, 2010). This could be problematic when considering organisms with complex life cycles. For example, holometabolous insects, in addition to embryonic development, undergo a complete metamorphosis (from pupa to adult), a process that could be considered a second embryonic development.

Currently, the most common definition of development refers to the set of processes through which an egg is transformed into an adult (Horder, 2010; Minelli, 2011). Already in 1880, Ernst Haeckel defined development in similar terms: "individual development, or the ontogenesis of every single organism, from the egg to the complete form is nothing but a growth attended by a series of diverging and progressive changes" (Haeckel, 1880).

Some authors criticize this egg-to-adult view to be an "adultocentric" view of development, and suggest instead to consider within the boundaries of development the whole life cycle of an organism (Gilbert, 2011; Minelli, 2011). Julian S. Huxley and Gavin R. de Beer said that development "is not merely an affair of early stages; it continues, though usually at a diminishing rate, throughout life" (Huxley and De Beer, 1963).

There have been recent attempts to construct a broader concept of development (Griesemer, 2014; Moczek, 2014; Pradeu, 2014) For example, Armin P. Moczek defines development as "the sum of all processes and interacting components that are required to allow organismal form and function, on all levels of biological organization, to come into being" (Moczek, 2014). The main challenge on adopting a new concept of development which is more inclusive, is to maintain its intuitiveness and applicability in scientific research.

Throughout this dissertation I will use the "common view" of development (Minelli, 2014), that considers the egg and the adult as the start and end of individual development respectively. However, and mainly for practical reasons, the major part of the analyses presented here (sudies I-III) are restricted to embryonic development.

1.1 Complexity

"The embryo in the course of development generally rises in organisation (...) I am aware that it is hardly possible to define clearly what is meant by the organisation being higher or lower. But no one probably will dispute that the butterfly is higher than the caterpillar."

Charles Darwin 1859

In this section, I will talk about the increase in complexity during embryonic development. Is important to mention that there is no consensus in the definition of

complexity, or how to measure it. It is indeed hard to find a definition of complexity that could be applied to the many different phenomena. A common intuitive notion of complexity relates to a system composed of many elements in a system and multiple interactions between these elements. However, some could see something as complex that is considered simple to others. Therefore, in this work I will use three different measures that relate to different intuitive aspects of complexity during embryonic development. **REVISAR** I will first try to explain different existing definitions of complexity applied to organisms and why I preferentially selected one and no another. Then, I will explore the relationship between complexity in evolution and development, and discuss the possibility of a trend in terms of complexity increase.

1.1.1 Different definitions of complexity

Complexity in informational terms

The use of informational terms (e.g., transcription, translation and code) in biology are widespread, specially in molecular biology (Smith, 2000; Yockey, 2005) More than just the use of informational terms in biology, information theory concepts like Shannon's entropy and mutual information have been used as a proxy to measure complexity. In the following paragraphs, I will briefly describe briefly these concepts and provide some examples of their use to address biological complexity.

Shannon's entropy Shannon's entropy is a measure of uncertainty. Given a set of n possible events whose probabilities of occurrence are $p_1, p_2 ..., p_n$, Shannon's entropy (H) can be defined as:

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

For a given n, the maximum H is equal to $\log n$ when all the events have the same probability (i.e., $\frac{1}{n}$) (Shannon, 1948). The logarithmic base of 2 corresponds to binary digits units, or *bits*.

This can be illustrated measuring H for a nucleotide position in the DNA sequence. As in principle, each DNA site can take four possible values (A, T, G or C), its maximal entropy can be calculated as:

$$H_{max} = -\sum_{i=A,T,G,C} p(i) \log_2 p(i) = \log_2 4 = 2bits$$

Cristoph Adami have used Shannon's entropy to define a "physical complexity" measure, that refers to the "amount of information that is stored in that sequence about a particular environment" (Adami, 2002). More specifically, Adami's complexity measure compares the maximum entropy of a specific DNA sequence with the "actual" entropy based on the actual probabilities $p_j(i)$ for each position j in the sequence. Given a pool of N sequences, $p_j(i)$ is estimated by counting the number $n_j(i)$ of occurrences of nucleotide i at position j, so that $p_j(i) = n_j(i)/N$ (for all positions $j = 1, ..., L$ of the sequence with length L) (Adami et al., 2000). The information content of a DNA sequence is then $I = H_{max} - H$ where:

$$H = -\sum_{j=1}^{L} \sum_{i=A,T,G,C} p_j(i) \log_2 p_j(i)$$

Adami assumes that if a sequence has not been under selective pressures each position in the sequence would have any of the four nucleotides with the same probabilities, so the actual entropy would be equal to the maximal, and consequently the information would be zero (Adami, 2002). He also considers that the "physical complexity" would serve as a good predictor of functional complexity (Adami, 2004). His information measure is related to the degree of conservation of a given sequence, which in the case of protein sequence has indeed been used to identify its functionality (Casari et al., 1995; Kellis et al., 2003; Hannenhalli and Russell, 2000).

Mutual information A concept related to Shannon's entropy that has been used in biological sciences is the concept of mutual information. Mutual information is a measure of the information in one variable about another. It is measured using the "conditional entropy" concept (the entropy of a variable Y given that X is known) introduced also by Shannon (1948). The mutual information $I(X;Y)$ of variables X and Y can be expressed as:

$$I(X;Y) = H(X) - H(X|Y)$$

where $H(X)$ is the Shannon's entropy of X and $H(X|Y)$ is the conditional entropy of X given Y. As Shannon's entropy, the conditional entropy is a measure of uncertainty. In this case, $H(X|Y)$ measures how uncertain we are of Y on the average when we know X (Shannon, 1948).

Therefore, mutual information measures how much information of one variable is contained in the other. It can also be though as a similarity measure (Yockey, 2005), as if X is identical to Y, the information of knowing X determines the value of Y. Some decades ago, there were great expectations on the use of informational measures to predict some features of an organism based on its DNA or protein sequences. For example, it was thought that the DNA sequence of a coding gene would determine not only the sequence of a protein, but also its 3D folded structure (Anfinsen, 1973). However, it is nowadays clear that other factors like post-translational modifications and the physico-chemical environment of the protein affect its structure (Kang and Kini, 2009) and that the DNA sequence is not sufficient to predict it.

This lack of correspondence between DNA and proteins have restrict the application of the mutual information as a similarity measure that compares DNA (Lichtenstein et al., 2015) or protein sequences (Gloor et al., 2005) separately.

However it has been proved useful to analyse different aspects of these molecular sequences, the informational approach has not been successfully applied to higher organisation levels (i.e., cells, tissues, organism) (Longo et al., 2012).

Algorithmic complexity Another definition of complexity that has been very popular is the algorithmic complexity. The algorithmic complexity (also called Kolmogorov complexity) of a data string would be the shortest algorithm necessary to describe such string. As the description of a string can be thought as a program to produce the data

(Wolfram, 2002) **BUSCAR UNA REFERENCIA MAS ANTIGUA**, the algorithmic complexity can also be defined as the shortest program that can produce the data.

Algorithmic complexity is related to randomness, as in algorithmic information theory if a program is as long as the data itself, then the data is considered to be algorithmically random (Wolfram, 2002). This measure of complexity that seems specially suited to analyse data strings is not easily applicable to other issues, like measuring phenotypic complexity. Even in the case of strings, it has been said that it is impossible to distinguish if most long sequences are random or not (Wolfram, 2002).

Also, it has been proposed that it is impossible to calculate the program-size complexity of anything, as it is impossible to prove that certain program is the shortest to produce some object (it would be only possible to prove upper bounds in its complexity if a program that produce the desired output is found; Chaitin, 1999). Some other authors have also said that algorithmic complexity, in which randomness is equated to high complexity, does not correspond to an intuitive notion of biological complexity (Adami, 2002), as living organisms are expected to show organization or order, far from randomness.

Spatial information There are also information-based measures applied to space. For example, Michael Batty (1974) used Shannon's definition of information and applied it to "spatial systems". He was specially interested in applying his measure to spatial systems such as a cities (Batty is a geographer). A brief description of Batty's method follows, for a full description see (Batty, 1974; Batty et al., 2014). For a certain location i, with a population P_i, the probability p_i is defined as the proportion of the population in i: $p_i = P_i/P$. To extend this to a probability density, the p_i is divided by the space available for the population, which is Δx_i. So the spatial entropy formula becomes:

$$S = -\sum_i p_i \log \frac{p_i}{\Delta x_i}$$

With this measure, the spatial entropy (and therefore spatial complexity) will reach its maximum when the probability is uniform $p_i = n$, and the distribution of land X in the different locations is also uniform $\Delta x_i = \sum_i \Delta x_i / n$. One application of this measure is to calculate if the entropy of the distribution density in a city has increased over time (Batty, 2010).

Another measure of spatial complexity based on information theory that has been proposed is the "spatial joint information" (Salazar-Ciudad et al., 2001) which indicates the relative entropy of a 1D arrangement, between cells with different "states" (based on the expression level of some gene). Salazar-Ciudad et al. (2001), used this measure to estimate the complexity of gene expression patterns in 1D cell arrangements produced with model genetic networks.

Computational metaphors

Many authors have used computational analogies to define development (Apter and Wolpert, 1965; Monod, 1963; Mayr, 1997; Davidson, 2001). Eric H. Davidson used the gene regulatory network (GRN) concept and a computational metaphor to explain development (and evolution). A GRN consists of DNA cis-regulatory elements, i.e.,

the regions in the vicinity of each gene which contain the specific sequence motifs at which those regulatory proteins which affect its expression bind; plus the set of genes which encode these specific regulatory proteins (i.e., transcription factors) (Davidson, 2001). For Davidson, development is then the outcome of spatial and temporal series of differential gene expression, that is controlled by a "regulatory program" (the GRN) built into the DNA.

A computational program, that is part of a computer system, contains a set of instructions that performs a specific task. The computer program needs a hardware, the set of physical objects that compose the computational system and where the computational program can be stored and execute. If the cell is considered as a computer system with the GRN as the computer program, then the hardware would be all the components of the cell including the genomic and cell structure and all the molecules present in the cell. However, in contrast to a computer system, the separation between the program and the hardware is not clear in a cell. The "genetic program" is affected by the components present in the cell ("hardware"), which in turn changes depending on the program (Oyama, 2000; Jaeger and Sharpe, 2014). For example, it is acknowledged that a cell might elicit different responses after the binding of an extracellular growth factor to a receptor in its membrane, depending the presence or relative abundance of key signal transducer molecules (Dailey et al., 2005). In other words, it can be said that the set of gene products present a cell define its "state" (Forgacs and Newman, 2005), and depending on the current state of a cell (which is in turn the product of the previous cell state plus extracellular signal), it will respond differently to a specific signal. This is the case from the very early stages of development, as the zygote transcription is regulated by the gene products that were maternally deposited. Also, it is important to mention that cells not only change their state during development, they also change their spatial distribution. A change in the spatial distribution (at a specific time during development) might have an affect in the outcome of development process (Salazar-Ciudad, 2010), for example, if some cells migrate or invaginate while producing a growth factor ligand, the cells that will receive the signal will be different. Therefore, it is cleat that not all the information necessary for the development is contained in the genome or GRNs.

Using again the computational metaphor, Davidson considered that these programs of gene expression, which are "installed and executed" as the embryo develops, could serve as a metric of complexity (Davidson, 2001). For illustrating his point Davidson describes an imaginary example of a GRN that increases its complexity in evolution: first, there is a set of downstream genes activated by a small network of TFs (each of them with only one *cis*-regulatory element), which in turn is controlled by a single upstream TF; then, TFs of the network gain cis-regulatory elements (so the circuitry is more intricate) and newly recruited intermediate regulatory TFs activate a different set of down-stream genes. Otherwise, the initial set of downstream genes is still controlled by the single upstream TF (Davidson, 2001).

Thus, in Davidson example, a small hierarchical network changes so that an additional layer is gained (intermediate TF) and the topology of the network changes: instead of one outcome (the initial set of downstream genes), now two outcomes are possible (with the additional set of downstream genes activated by the new intermediate TF).

McShea's view of complexity

Daniel W. McShea has provided useful definitions of biological complexity. One of his definitions is: "complexity of an organism is the amount of differentiation among its parts or, where variation is discontinuous, the number of part types" (McShea, 1996, 2015). This definition can be used at different hierarchical levels of biological organization, e.g., tissues, cells, genes. Indeed, a measure of morphological complexity that has been favoured by some authors is the number of cell types that compose an organism (Valentine et al., 1994; Bell and Mooers, 1997; Bonner, 2004). This definition of complexity is not exempt of complications, as there is no clear criteria of how to define a cell type or how to determine, during development, when a new cell type has formed.

Importantly, with this definition (complexity as the number of parts), the complexity at different levels are not necessarily correlated. This lack of correspondence at different levels using "the number of part types" its evident comparing the number of genes with the number of cell types. Before the release of the first eukaryotic genome sequences, it was expected that the number of genes would correlate with an intuitive perception of organismal complexity, ranking complexity as yeast < nematodes < flies < humans (Hahn and Wray, 2002) (this intuitive notion of complexity correlates with the number of cell types in metazoans; Valentine et al., 1994). However, this expectation was proved to be wrong and this lack of correlation between "intuitive complexity" and genes number was called the "G-value paradox" (Hahn and Wray, 2002). Before that, the lack of correspondence between genome size and organism complexity (using again an intuitive notion of complexity), or "C-value paradox", was also noted. The lack of correspondence between the number of genes with an intuitive notion of complexity is now partly explain by some authors by the amount of post-transcriptional regulation. It has been therefore propose that alternative splicing (**REF**) or the amount of miRNA regulation is related with the complexity in vertebrates (**REF**). **PARTE DE LA PARADOJA SE RESUELVE CON EL HECHO DE COMO SE REGULA EL DESARROLLO EN LOS ORGANISMOS, QUE NO ES DEPENDIENTE SOLAMENTE DEL NUMERO DE GENES, POR EJEMPLO LA TOPOLOGIA DE REDES ES UN MECHANISMO QUE AFECTA LA COMPLEJIDAD FENOTIPICA**

McShea also makes the distinction between "object complexity" that refers to the number of parts of a system and "process complexity" that refers to the interaction among parts in a system (McShea, 1996). This could be illustrated with the number genes (object complexity) and the number of gene-gene interactions (process complexity). Gene-gene interactions would refer to the regulation of a gene expression by the binding of another gene product (transcription factor) to its promoter region. Using this definition, two different organisms would have the same object complexity if they have the same number of genes, but one would have a higher process complexity it has more gene-gene interactions than the other.

1.1.2 Complexity Increase in Development and Evolution

The increase in complexity in an organism during embryogenesis is probably one of the most intuitive processes of animal development.

It is commonly seen even as one of its defining characteristics. Eric H. Davidson described the progressive increase in complexity as the "essence" of development (Davidson, 2001). Despite of the widely accepted view of complexity increase in development, there

is no consensus of how to define it, much less on how to quantify it (Oyama, 2000).

Using the number of cell types, the increase of complexity during development is self-evident: in vertebrates, the embryo begins with one cell type (the zygote) and concludes with more than 200 cell types (Alberts et al., 1994).

On the relationship between the increase of complexity in Evolution and Development

The connection between the increase in complexity during development and evolutionary time has been largely discussed. Haeckel was one of the first who made explicit hypotheses about the connection between the development and evolutionary patterns in his "Biogenetic Law" (see Box 1). These laws imply that the increase in complexity we see during development is a reflection of a similar increase in complexity that has occurred through evolution. Early views of evolution saw the increase in complexity as inexorable, with all the species descending from simpler ancestral forms (Lamarck, 1809; Haeckel, 1874), and with the human species as the latest and more perfect product of the evolution of animals (Haeckel, 1874).

Recent views recognize that complexity of can increase or decrease in a lineage. Using the number of cell-types as complexity measure, there are clear examples of taxa that have decreased their complexity over time, specially in parasites (Canning and Okamura, 2003; Arthur, 2010) (although morphological simplification can not be considered universal in parasitic taxa Poulin, 2011).In the other hand, there are many lineages that have remain unicellular (therefore their complexity would have remained constant), while some lineages (e.g., vertebrates) have increased their complexity. Therefore, it seems that there is no unique trend to increase the complexity over time. Therefore, the complexity of a specific lineage might decrease, increase or stay the same (see Figure **??**.

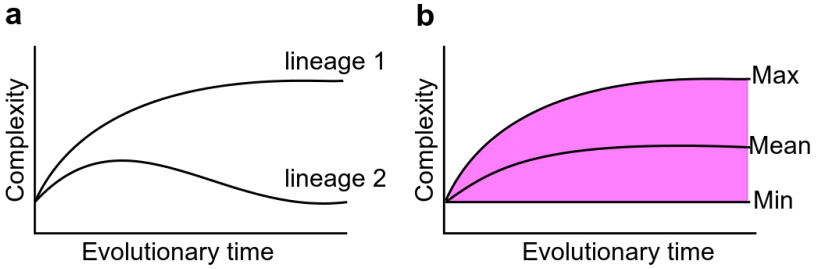

Figure 1.1: Two lineages with different complexity change through their evolutionary trajectories. b) Representation of the minimum, mean and maximum complexity of many lineages over evolutionary time in which the minimum stay constant while the mean and maximum increase. Redrawn from (Arthur, 2010).

If we consider uni-cellularity as the minimum complexity, it can be said that complexity has remained more or less constant in evolution, as unicellular organisms like bacteria, have been present from 3.5 billion years. However, if we consider instead maximum complexity a trend for increasing complexity would be apparent, as complexity would have increased with the appearance of simple multicellular organisms (only few cell types) and would have further increase until the appearance of organisms composed of hundreds of cell types. Some authors consider that this apparent trend of increasing complexity is the product of natural selection (**REF**). This notion can lead to circular reasoning, saying that organisms are complex because they have been selected for that,

and that they have been selected because they are complex. **EXPLICAR MEJOR**

Other authors consider that this apparent trend does not necessarily imply that it has been selected for (McShea, 2015), and that this trend might appear even in scenarios without natural selection. For example, if we consider an evolutionary scenario in which the complexity of the organisms follow a random walk scenario without any selection regime but with the condition of a lower boundary (i.e., is not possible to have less than once cell), and starting from unicellular organisms, we will see an increase of the maximum complexity. with an initial increase in the mean complexity (in a random walk the mean distance of a point after n time steps is $\sqrt[2]{n}$) **DESCRIBIRLO MAS CLARAMENTE**

Compartmentalization in development

The notion of an increase in complexity during embryogenesis is tightly related to the concept of embryo compartmentalization. In here, I will refer to compartmentalization as the subdivision of the embryo in different parts (or compartments) during development. How the different parts of the embryo can be defined based on the cell-phenotype or gene expression profile.

It is usually considered that the earliest compartments that are formed in the embryo define the main body axis, i.e., the anterior/posterior (A/P) and dorsal/ventral (D/V) axes. Later on, smaller compartments of the embryo would be formed, e.g, limbs, eyes or internal organs. In this manner as development proceeds, it is expected that spatial compartments would be progressively specified at an increasing finer resolution (Davidson, 2001). Furthermore, the increasing compartmentalization of the embryo during development can be conceptualized as the progressive spatial restriction of gene expression to subsequently smaller regions in the embryo. Sean Carroll defines this process(Carroll et al., 2001) as:

i. In early development, genes have a broad expression in the embryo and define the main axes of the body.
ii. Later, genes define smaller compartments like organs and appendages (field-specific selector genes).
iii. Finally, genes become expressed in specific cell types like muscle and neural cells (cell-type specific selector genes).

It is important to note that this would imply that, in general, the area of expression of a gene in the embryo would decrease during development (relative to the area of the whole embryo).

A well known definition of developmental compartment was proposed by García-Bellido et al. (1973). They defined compartments as differentiated populations of cells (at the gene expression level) that do not intermix between them that are formed from initially homogeneous contiguous cells. The definition used in here is related to the one of Garcia Bellído et al., but in contrast to it, does not rely on the identification of a boundary formation between different cell populations that would prevent cell mixing between them.

Complexity at the molecular level

For some authors, the increase in complexity in an organism during development (reflected by the increase of number of cell types), should be associated with an underlying

complexity at the molecular level (Davidson, 2001; Arthur, 2010), following the reasoning that:

i. In development, complexity increases with time as new cell types form.
ii. Different cell-types are characterized by the differential expression of genes.
iii. Therefore, a complex organism (composed of many cell-types) has to contain a complex gene expression regulatory machinery to produce the different combinations of expressed genes in each cell-type (Davidson, 2001).

Gene expression regulation It is widely acknowledged that the spatio-temporal regulation of gene expression in development is crucial for the progressive compartmentalization of the embryo. More than fifty years ago, Jacques Monod and François Jacob (Jacob and Monod, 1961) published in a seminal work a model of the genetic regulatory mechanism in bacteria. The most important conclusion of this paper was the existence of "regulator" genes that control the production rate of proteins from "structural" genes, and that mutations in "regulator" genes affect the regulatory mechanism but not the structure of the regulated protein. In the same paper they suggested that these regulator genes may affect the synthesis of several different proteins (Jacob and Monod, 1961).

Nowadays the process of gene activation is known in great detail. The "regulator genes" Jacob and Monod studied are transcription factors, proteins that bind to DNA to promote or repress the transcription of a gene.

This transcriptional regulation represents however only level of gene expression regulation. There are many other mechanisms that regulate the production of gene products. These include 3' untranslated regions (UTR) (Grzybowska et al., 2001), small interference RNAs (siRNAs) (Filipowicz et al., 2005), translational (Kozak, 1992; Kapp and Lorsch, 2004) and post-translational (Mann and Jensen, 2003) regulation of gene expression.

At least two regulatory levels have been suggested to have a causal role in the increase in complexity in different lineages: transcriptional regulatory level (as mentioned above) (Davidson, 2001) and miRNAs. The role of miRNAs (non-coding RNA molecules that negatively regulate gene expression) was proposed after the observation that miRNAs are found only in protostomes and deuterostomes and not in sponges or cnidarians, and that they are specifically expressed in certain cell-types, tissues or organs (Sempere et al., 2006). It could be expected however that the complexity of an organism could be reflected at any level of gene expression regulation, whether transcriptional, post-transcriptional, translational or post-translational.

Additionally to all these regulatory levels that occur within a cell, the spatio-temporal regulation of gene expression depends on signalling between neighbouring cells and morphogenetic events (Gilbert, 2014). Importantly, cell-cell signalling can affect the transcriptional state of a cell (through a signalling pathway). Additionally, the specific response of a signalling event is dependent on the transcriptional state of the cell.

1.1.3 Spatial complexity

Until this point, I have focused on the concept of compartmentalization as one aspect that reflects the increase in complexity during development. Another aspect that is intuitively related to the increasing embryonic complexity is the shape (or form) of the embryo. Focusing on the shape of the embryo, embryonic development can be thought

as a process that starts with a simple spherical or oblate fertilized egg and that ends with complex shapes and forms (in the adult or larva) (Forgacs and Newman, 2005).

In addition to the external shape of the embryo, the shape complexity of its internal structures (e.g., organs) is expected to increase during development (Sharpe, 2003). However, is not always possible to appreciate the morphological change of inner structures at simple view. The first attempts to describe the shape of internal organs during the development in vertebrates was in the 19th century, and it required section cutting and wax reconstruction (Hopwood). The use of staining techniques have facilitated the visualization of the inner morphology of the embryo. Techninques such as the horseradish peroxidase staining facilitated not only the visualization of the inner morphology, they were also crucial to create the first fate maps, while being able to trace the cell-lineage of different organs.

Since now is known that many genes are expressed in a tissue/cell type specific manner, another useful tecnique to visualize inner structures is the use of labelling techniques that highlight the distribution of such tissue-specific gene products (e.g., whole-mount in situ mRNA hybridization or immuno-histochemistry techniques). **GENE EXPLICA EN PART DISTRIBUCION Y FORMA DE ORGANOS AUNQUE NO SEA UNA CORRESPONDENCIA DE UNO A UNO.**

If the embryo external and internal morphology are expected to increase their spatial complexity, and the gene expression patterns (visualized with a labelling technique) is expected to correspond to specific regions (e.g., internal organs) or the whole embryo (in case the expression is ubiquitous), therefore the shape of gene expression patterns can be used to describe the morphological complexity of the embryo. Also, it could be expected that when analysing a large number of genes, the shape complexity would increase at a statistical level.

The study of morphometrics refers to the quantitative analysis of morphological shape. In the last decades, morphometric tools have been used widely as a tool to quantify, characterize and compare biological shapes (James Rohlf and Marcus, 1993). In the next paragraphs I will briefly explain the most important morphometric methods. For an extensive review, see (Bookstein, 1997; I. L. Dryden, 1998; Zelditch et al., 2008; Slice, 2005).

Morphometrics

In morphometrics, shape refers to the geometric properties of an object that are invariant to location, scale and orientation (Slice, 2005). Many of the modern morphometric analyses are based on the use of "landmarks", which refer to precisely located points that establish a clear one-to-one correspondence between the samples under study (Klingenberg, 2010). To extract only the shape information from the landmarks, the variation in size, position and orientation are usually removed with a technique called "Procrustes superimposition" (I. L. Dryden, 1998). Although there are many different morphometric methods, they can be divided in four main categories: "traditional morphometrics", "geometric morphometrics", outline analysis and surface analysis (Slice, 2005).

The **"traditional methods"** refer to the application of multivariate statistics, like Principal Component Analyses (PCA), to the direct measurement of lengths, widths or ratios of specific structures. Some typical applications of these methods are the classification of species or sexes (Jolicoeur and Mosimann, 1960) using lengths, widths or angles between landmarks (I. L. Dryden, 1998).

Geometric morphometrics analyses use instead geometric coordinates of morphological landmarks (Mitteroecker and Gunz, 2009; Zelditch et al., 2012). As with the traditional methods, PCA can be used for analysing the shape variation in the dataset (Klingenberg, 2010). A variant of landmark analyses is the use of "semi-landmarks". Semi-landmarks are equally spaced points around an outline, usually between "real landmarks". Semi-landmarks are therefore used when only a few landmarks are recognisable. For example, in the analysis of the shape of hands, "real landmarks" can be placed in the tip of the fingers, and the semi-landmarks would be placed along the hand outline. After recording the semi-landmarks, procrustes superimposition and multivariate analyses can be used as with ordinary landmarks (I. L. Dryden, 1998).

Outline analyses are specially relevant when it is not possible to identify comparable landmarks between samples. One type of outline analyses is the Elliptic Fourier description (Kuhl and Giardina, 1982), which uses an orthogonal decomposition of a curve into a sum of harmonically related ellipses (Ferson et al., 1985). The extracted harmonics can then be analysed with PCA. A classical example of the Fourier description is the analysis of mussel shells (which can be represented as a closed outline) done by Ferson et al., 1985. Another outline method is the Eigenshape analysis, which uses outline coordinates to calculate angles between points to provide a map around the outline. More specifically, shape is represented as the shape function $\phi^*(l)$, the normalized net angular change in direction ϕ at each step around the perimeter (l) (the normalization can be based on the deviation from a circle, or from the sample mean) (Lohmann, 1983). Then, the major directions of observed and measured shape variation is analysed by means of eigenanalysis. Eigenshape analysis (a type of Principal Component Analysis) derives a set of empirical orthogonal shape functions by an eigenfunction or PCA of a matrix of correlation between shapes (Lohmann, 1983).

Surface analyses are used when comparing 3D objects (usually represented as Cartesian coordinates x,y and z of points on the object's surface) with limited landmark information. For example, a vertebrate skull has many identifiable anatomical landmarks in the face but only a few can be defined unambiguously on the smooth braincase (Mitteroecker and Gunz, 2009). In order to deal with this, Gunz et al., 2005 extended the use of 2D outline semi-landmarks to 3D surfaces. 3D semi-landmark methods are based on allowing the points to "slide" along the surface until some measure of shape difference (e.g., bending energy of a thin-plate spline) among the configurations is minimized (Mitteroecker and Gunz, 2009).

Topographic anaytical methods

Another approach to 3D surfaces is the use of topographic analytical methods. These methods, which apply concepts from Geographic Information Systems (GIS), have been use to quantify teeth surfaces as if they were landscapes (Jernvall and Selänne, 1999; Winchester et al., 2014).

New topographic analytical methods that do not rely on GIS have been recently developed. This is the case of the Dirichlet normal energy (DNE), a method for quantifying surface bending using concepts from differential geometry (Bunn et al., 2011). This method quantifies the deviation of a surface mesh from being planar (Bunn et al., 2011). A brief explanation of the DNE follows, for a complete description see (Bunn et al., 2011; Winchester, 2016). For each polygon in the surface mesh, DNE calculates its energy value $e(p)$. The energy value quantifies change in the normal map around a

polygonal face. The DNE value of the whole surface is the sum of all the energy values $e(p)$ of the polygonal mesh surface. DNE values increase with both convexities and concavities on a surface.

1.2 Adaptation

In this section, I will start with the definition of the concepts of adaptation (although there are more than one definitions of adaptation; Endler, 1986), and natural selection. Then I will introduce some of the methods that are used to estimate adaptation, with a focus on molecular methods.

On the concept of adaptation

Usually adaptation refer to two different things, to an adaptive trait or to the process to become adapted (Endler, 1986). George Gaylord Simpson (1953) defined adaptation in the following way:

"*an* adaptation is a characteristic of an organism advantageous to it or to the conspecific group in which it lives, while adaptation or the process of adaptation is the acquisition within a population of such individual adaptation" (italics by the author)

An adaptation (i.e., an adaptive trait) is usually related to a specific function of the organism. For example, the beak variations (in size, width and depth) in the Darwin's Galapagos finches, a classic example of adaptive traits, are related to the alimentary function of the finches, so that each species is specialized in a specific diet. The notion of adaptation existed before Darwin, but since Darwin it is closely related to the concept of natural selection. Under the current evolutionary framework, an adaptation, arising due to intrinsic natural variation, will be fixed in a population by natural selection due to the advantage it confers to their bearer organisms.

Natural selection

Charles Darwin, in its 1859's *Origin*, defined Natural selection as follows:

> *" Owing to this struggle (for life), variations, however slight and from whatever cause proceeding, if they be in any degree profitable to the individuals of a species (...) will tend to the preservation of such individuals, and will generally be inherited by the offspring. The offspring, also, will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive. I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection"* (Darwin, 1859).

More recently, and following the Darwininan concept of natural selection, Jhon A. Endler (1986), defined it as a process in which, given that a population has:

a. **variation** among individuals in some attribute or trait;
b. **fitness differences** (consistent relationship between that trait and mating ability, fertilizing ability, fertility, fecundity, and, or, survivorship);
c. **inheritance** (consistent relationship, for that trait, between parents and their offspring, which is at least partially independent of common environmental effects).

Then:

i. the trait frequency distribution will differ among age classes or life-history stages, beyond that expected from ontogeny;
ii. if the population is not at equilibrium, then the trait distribution of all offspring in the population will be predictably different from that of all parents, beyond that expected from conditions a and c alone.

Conditions a, b, and c are necessary and sufficient for the process of natural selection to occur, and these lead to deductions i and ii (Endler, 1986).

Condition a relates to phenotypic changes across generations. Importantly, phenotypic changes, whether new characters or modifications of existing characters in the adult/larva, are produced from changes in development. For example, the difference in the beak size and shape between the famous Galapagos Darwin's finches, a classic example of adaptive change under natural selection, has been shown to be regulated by the differential expression of the genes CaM and BMP4 during development (Abzhanov et al., 2006).

Therefore, even when natural selection acts in the adult/larva phenotype, the changes that lead to an adaptation should be traceable during the development of such trait.

Methods to detect natural selection

There are many different methods designed to detect natural selection in natural populations. Jhon A. Endler classified ten different methods with diverse ability to detect natural selection (Endler, 1986). Some of these methods test directly the conditions (b and c) required by natural selection, while others test the predicted outcome of natural selection in a population.

There are many studies that have aimed to detect natural selection directly in the phenotype. Usually, these studies are based on the estimation of selection gradients of a quantitative trait (a measurable phenotype that usually depends on the cumulative actions of many genes and the environment). A selection gradient of a trait refers to the relation of the trait values and fitness. Is calculated as the slope of a linear regression of fitness on the trait value (Barton et al., 2007). Most of these studies are based on the analysis a single trait or a small number of traits of an organism (Hoekstra et al., 2001; Hereford et al., 2004) which are usually selected already under the suspicion to be adaptive. On the contrary, there is practically no study that has attempted to estimate natural selection over the entire organism.

Among the methods that test the predicted outcome of natural selection we find the molecular methods. The molecular methods are based on the assumption that given that changes leading to an adaptation are (at least partially) caused by DNA mutations, the effects of natural selection could be traceable looking at the DNA sequence. There is an entire field within evolutionary biology, namely molecular evolution, dedicated to explain the sequence changes in molecules as DNA, RNA and proteins in evolution

In the next sections, due to its relevance in this work I will only focus on the molecular methods to detect natural selection.

1.2.1 Molecular evolution

The theoretical basis of the molecular evolution field includes concepts from evolutionary biology and population genetics. At the DNA level, any transmissible change in the

sequence is considered a mutation. The most simple change is a point mutation, also called single nucleotide polymorphism (SNP), which is a change in a single nucleotide in the DNA sequence of a locus in an individual.

Variation at a particular DNA site within the individuals of a species or population is referred as polymorphism, while divergence refers to variation at a specific DNA site in individuals from different species.

SNPs occur in non-coding and coding DNA sequences. A SNP that occurs in a coding sequence is classified in two categories, depending on its effect on the protein sequence: i) synonymous mutation and ii) non-synonymous mutation. A synonymous mutation does not affect the amino-acid sequence of the protein (although it can affect its function (Kimchi-Sarfaty et al., 2007) or the gene transcriptional efficiency (Xia, 1996). A non-synonymous mutation affects the amino-acid sequence of the protein whether by changing a single amino-acid (missense mutation) or by producing a stop codon (non-sense mutation) which results in a truncated version of the protein.

As the non-synonymous mutations can affect dramatically the structure and function of the protein, it would be expected that most non-synonymous mutations have a negative fitness effect. However, it is also expected that a fraction of non-synonymous mutations, or adaptive substitutions, would have a positive fitness effect that (depending on the strength of the fitness effect and several population genetics parameters) could lead to the fixation of that mutation in the population.

An important branch of the molecular evolution field is dedicated to the identification of adaptive substitutions in a species, which has lead to the development of many statistical tests. Importantly, these tests are based on the neutral theory of evolution, proposed by Kimura (Kimura, 1968).

Neutral theory of evolution

In 1968, Mooto Kimura calculated the average rate of nucleotide substitutions in the evolutionary history of mammals. The result of his calculations was that, on average, one nucleotide has been substituted every 2 years. For him, this very high rate of substitution was only explainable if most mutations were almost neutral in natural selection (Kimura, 1968). which was in contrast with the prevailing view at the time that practically no mutations were neutral.

In 1969, Kimura proposed that the majority of amino acid substitutions that occurred in proteins are the result of random fixation of selectively neutral or nearly neutral mutations (Kimura, 1969). In the same year, King and Jukes (King and Jukes, 1969) proposed independently practically the same hypothesis.

Two important assumptions of the neutral theory of molecular evolution were:

i. Deleterious and adaptive mutations are rapidly purged and fixed in a population respectively.
ii. Polymorphism is a transitory phase between random fixation or extinction due to genetic drift.

Importantly, the neutral theory provided a set of testable predictions, providing a null-hypothesis for adaptive molecular evolution.

From neutral to nearly neutral theories

In the subsequent decades after the proposal of the neutral theory of molecular evolution, much more protein sequence data became available, which made evident the great variation in the evolution rate of proteins. To account for this, Kimura and Ohta stated that "functionally less important molecules or parts of a molecule evolve faster than more important ones" (Kimura and Ohta, 1974). Then, Ohta propose that slightly deleterious mutations might be common in amino acid substitutions (OHTA, 1973). Later, it was proposed that half of the protein substitutions would be advantageous and the other half deleterious (Gillespie, 1994). Therefore, the neutral model was replaced by a nearly neutral model with only deleterious substitutions, which in turn was replaced by one with a mixture of positive and negative effects (Ohta and Gillespie, 1996).

At the end of the 1970's comparative analyses of protein sequence data began to be replaced for analyses on DNA sequence data, which revealed that synonymous substitutions within coding regions are more frequent than non synonymous (those that change an amino acid) substitutions. From the early 1990s, the expectations of the nearly neutral theory at the DNA sequence level are that substitutions in non coding DNA and synonymous substitutions in coding regions are neutral and amino acid substitutions can be deleterious, neutral or advantageous (Ohta and Gillespie, 1996). Statistical methods were then devised to test such expectations.

1.2.2 Estimating adaptation at the molecular level

One of the most popular tests to estimate adaptation at the molecular level has been the McDonald-Kreitman test (MKT), which is used to detect adaptive substitutions comparing the relative numbers of synonymous and non-synonymous differences within a species with those numbers between closely related species.

McDonald-Kreitman test

John H. McDonald and Martin Kreitman developed this test in 1991 when analysing the divergence in the Alcohol dehydrogenase (Adh) locus in three *Drosophila* species (McDonald and Kreitman, 1991). The main assumption of the MKT is that the substitutions in a protein are neutral if the inter-specific ratio of non-synonymous (Dn) to synonymous (Ds) changes is equal to the intra-specific ratio of non-synonymous (Pn) to synonymous (Ps) changes (i.e. $Dn/Ds = Pn/Ps$). Any departure from this equality would imply the action of positive or negative selection. Importantly, MKT assumes for simplicity that non-synonymous mutations are either strongly deleterious, neutral or strongly advantageous (McDonald and Kreitman, 1991).

In other words, this method assumes that a protein in a given phylogeny has a specific substitution rate for synonymous and another for non-synonymous substitutions. Therefore, when comparing two proteins from two different species that have evolved under neutral conditions, the *total* number of each type of substitutions would depend on the time since the separation between species. If the proteins are from individuals from the same species, it would depend on the time elapsed since the separation of the within-species branches of the phylogeny. However, the ratio of non-synonymous to synonymous substitutions is expected to be the same in both inter-specific and intra-specific cases. In the case of non-synonymous mutations under positive selection (synonymous

mutations are expected to be always neutral), the equality of these ratios would disappear. As mutations under positive selection are expected to spread through a population rapidly (and are not expected to very common) they are not expected to be present as polymorphic (i.e., inter-specific) variation, but only as divergent (i.e., inter-specif) one.

Therefore, in the presence of mutations under positive selection, the ratio of non-synonymous to synonymous variation within species should be lower than the ratio of non-synonymous to synonymous variation between species (i.e. $Dn/Ds > Pn/Ps$; see Fig. 1.2). On the contrary, if the observed ratio of non-synonymous to synonymous variation between species is lower than the ratio of non-synonymous to synonymous variation within species (i.e., $Dn/Ds < Pn/Ps$) then negative selection is at work.

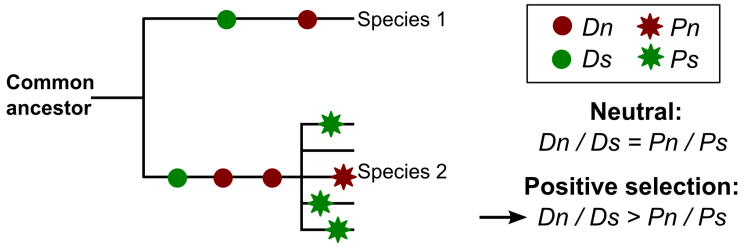

Figure 1.2: McDonald-Kreitman Test (MKT). MKT compares the ratio of non-synonymous (Dn; red circle) to synonymous (Ds; green circle) divergence (Dn/Ds) to the ratio of non-synonymous (Pn; red star) to synonymous (Ps; green star) polymorphic changes (Pn/Ps). Positive selection is detected when $Dn/Ds > Pn/Ps$, as in the example shown in the left.

Although the MKT has been proved robust to many sources of error (e.g., variation to mutation rate across the genome), it can be affected by the presence of slightly deleterious mutations or demography (Messer and Petrov, 2013; Eyre-Walker et al., 2006). The effect of slightly deleterious mutations are related to the effective population size (N_e). In a population with a low N_e, slightly deleterious mutations would have more probabilities of fixation by random genetic drift contributing more to polymorphism than to divergence, underestimating the proportion of adaptive changes (Messer and Petrov, 2013).

Recently, sophisticated methods based on the MKT have been developed to correct for underestimation of adaptive evolution in the presence of slightly deleterious mutations.

Distribution of Fitness Effects

Even when for simplicity the mutation effects are usually classified in strongly advantageous, neutral, and strongly deleterious, there is actually a continuum of selective effects, from strongly deleterious, to highly adaptive mutations, with weakly deleterious, neutral and slightly adaptive mutations in between (Eyre-Walker and Keightley, 2007).

The relative frequencies of all these types of mutations is called the Distribution of Fitness Effects (DFE). In order to know the DFE, a few experimental approaches exist. The most direct method is whether to induce (Sanjuán et al., 2004) or to collect (Mukai, 1964) spontaneous mutations and assay their effects (fitness) in the laboratory. As it can be expected, these experiments require many generations to gather sufficient data, so these approaches have been used mainly in micro-organisms (Eyre-Walker and

Keightley, 2007). A caveat of these experimental approaches is that, in order to identify the effect of a mutation, its effect has to be detectable in a fitness assay. In fitness assays however, only effects with relatively large effects are usually detected. Therefore, these methods give valuable information for mutations with relatively large effects.

An alternative approach is to infer the DFE by analysing patterns of DNA sequence differences at intra and inter-specific level (polymorphism and divergence respectively). The methods using this approach rely mainly on two assumptions:

i. the probability that a mutation spreads to a certain frequency in a population (or to fixation) depends on the strength of selection (positive or negative) acting on it. Severely deleterious mutations have lower probability to reach a high frequency in a population.

ii. the efficiency of selection depends on the effective population size. With a high effective population size, selection is more efficient and a smaller proportion of mutation will behave as effectively neutral.

The "absolute strength" of selection on a mutation is then measured as $N_e s$, the product of the effective population size (N_e) by the selection coefficient (s) of the mutation. Mutations with $N_e s$ much less than 1 are effectively neutral, while $N_e s$ greater than 100 have no chance to appear as polymorphism (Eyre-Walker and Keightley, 2007).

DFE-alpha method

Eyre-Walker and collaborators proposed a method to estimate both the DFE and the proportion of adaptive nucleotide substitutions (α) using polymorphism and divergence data (Eyre-Walker and Keightley, 2009). More specifically, they use the polymorphism site frequency spectrum (SFS) to estimate the DFE and then use this estimated DFE to estimate the proportion of substitutions under positive selection between species.

—- **how from SFS to DFE?**

This method, assumes that there are two types of nucleotide sites: i) sites at which all mutations are neutral and ii) sites at which some of the mutations are subject to selection (positive or negative). Also it is assumed that any new adaptive mutation in a population would not be detected in the polymorphic phase but only in the divergent one (as the advantageous mutations would fix rapidly in a population), and that the DFE can be represented with a gamma distribution (Figure 1.3).

The divergence at the neutral sites is then proportional to the mutation rate per site and the predicted divergence at the selected sites (in the absence of advantageous mutations) is proportional to the product of the mutation rate together with the average fixation probability of a selected mutation. This probability of fixation is is inferred based on the DFE and other parameters estimated from the polymorphism data analysis(Eyre-Walker and Keightley, 2009). The difference between the observed and predicted divergence therefore estimates the divergence due to adaptive substitutions. Using this method Eyre-Walker and collaborators estimated that approximately 50% of amino acid substitutions and approximately 20% of substitutions in introns are adaptive (Eyre-Walker and Keightley, 2009).

Messer and Petrov performed molecular evolution simulations to test if the estimates of different tests, like the MKT and the more sophisticated DFE-alpha, are accurate under different realistic gene-structure and selection scenarios (Messer and Petrov, 2013), specially in the presence of genetic draft (stochastic effects generated by recurrent selective sweeps at closely linked sites) and background selection (interference among linked

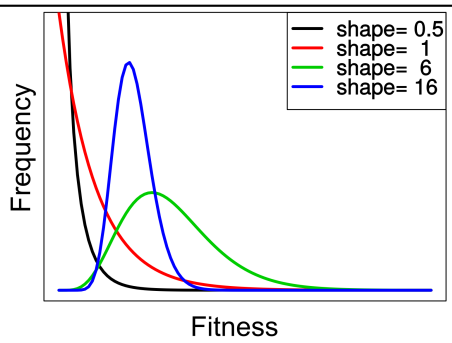

Figure 1.3: Example of different Distribution of Fitness Effects (DFE) represented by a gamma distribution. Many distributions can be represented by modifying the shape parameter of a gamma distribution, from a leptokurtic (shape parameter less than 1) to an exponential (shape parameter equal to 1) or a skewed normal distribution (shape greater than 1).

sites by lightly deleterious polymorphisms). They found that in the presence of slightly deleterious mutations, MKT estimates of α are severely underestimated and that DFE-alpha is very accurate to calculate α even in the presence of genetic draft, background selection or demography changes (Messer and Petrov, 2013).

1.2.3 The *Drosophila melanogaster* Genetic Reference Panel

The DGRP is a publicly available tool for molecular population genomics analyses. It consists of 192 inbred strains derived from a single outbred *Drosophila melanogaster* population. The inbred lines were constructed from collected mated females from a Raleigh (North Carolina, USA) population, followed by 20 generations of full-sibling inbreeding of their progeny (Mackay et al., 2012). 168 inbred lines were then sequenced using Illumina (129 lines), 454 sequencing (10 lines) or both (29 lines). Therefore, the DGRP contains a representative sample of naturally segregating genetic variation.

Mackay et al., 2012 used the DGRP sequence data in combination with genome data from *Drosophila simulans* and *Drosophila yakuba* to analyse polymorphism and divergence, the recombination landscape, and infer the action of natural selection on an unprecedented genome-wide scale. They found that the patterns of polymorphism differ by autosomal chromosome region, and between the X chromosome and autosomes, contrary to the divergence patterns. Using version of the MKT test, they estimated that on average 25% of the fixed sites between *D. melanogaster* and *D. yakuba* are adaptive (24% non-synonymous, 30% in introns and 7% in UTR sites) (Mackay et al., 2012).

1.3 Drosophila

The fruit fly, *Drosophila melanogaster*, has been a great valuable tool for biological research. Its use as a model system dates back to the beginning of the 20th century. In 1908, Thomas H. Morgan (see section X) started to grow flies in large quantities to study gene mutations. At that time, the gene concept was an abstract one, as the nature and location of the genes was still disputed. The main advantages of using flies were their rapid generation time, it was easy to culture and cheap to maintain (Arias,

2008). In his lab, at the University of Columbia, Morgan encountered a fly with white eyes (the wild-type eye color is red), which became a subject of his research for many years. Eventually, he discovered that the allele of the gene, that he called *white*, was located in a sex chromosome, demonstrating for the first time the sex-linkage of genes (Morgan, 1919). Morgan's students also demonstrated that mutations were inducible with X-rays and introduced the use of "balancer" chromosomes to keep stable stocks of mutants (Arias, 2008). The research carried in Morgan's lab laid the basis modern genetics, and its fly room became a central node in the genetics research, establishing *Drosophila* as a organism model.

However, *Drosophila*'s development was difficult to study, as the embryos were not large enough to experimentally manipulate them, and not transparent enough to visualize with a microscope (Gilbert, 2014).

In the next subsections, I will describe briefly the *D. melanogaster* life cycle with special focus on its embryonic development and the blastoderm fatemap and the relation of fate maps with gene expression maps.

1.3.1 *D. melanogaster* life cycle

Drosophila melanogaster is a holometabolous insect, which means that it goes through a complete metamorphosis, i.e., the larva and the adult forms are very different. The larva grows and passes through two moults before becoming a resting stage called a pupa in which the body is remoulded to form the adult.

Much of the adult body is formed from the imaginal discs and the abdominal histoblasts which are only present as undifferentiated buds in the larva.

Developmental stages

In Table 1, a brief summary of the embryonic development of *D. melanogaster* is shown. For a comprehensive lecture, see (Campos-Ortega and Hartenstein, 1985; Roberts, 1998; Gilbert, 2014).

The staging system shown in Table 1, correspond to the 16-stage system proposed by (Roberts, 1998), with approximate developmental timings at 22 °C. The numbering of the stages shown are similar to the one proposed by Campos-Ortega and Hartenstein (1985), except that the latter add a stage 17 to the fully differentiated embryo. The 16-stage system is used by the BDGP (Tomancak et al., 2002). Therefore, Table 1 can serve as a reference when mentioning specific developmental stages in this work.

1.3.2 Fate map

The "fate map" is a very important concept in developmental biology. Its name refers to the practice of cartography (or map making), i.e., contructing two-dimensional (2D) representations of a usually three-dimensional (3D) space. In the case of a fate map, the prospective fate is mapped onto the 2D representation of usually an early embryo (Gilbert, 2007). The first fate maps were constructed by tracking cell lineages to identify cell fate, only by observation. In 1905, Conklin tracked the cell lineage of the tunicate embryo, providing the first fate map (Conklin, 1905). In the 1930's Hörstaduius provided the fate map of the sea urchin embryo, also by cell lineage tracking. These two species

Table 1. Embryonic stages of *D. melanogaster* and morphological criteria for identifying approximate ages (from Roberts, 1998)

Stage	Developmental time	Morphologic features and main developmental events
1	0 to 15min	**Freshly laid egg.** Homogeneous cytoplasm
2	15min to 1h 20min	**Early cleavage.** A cap of clear cytoplasm becomes visible at the posterior pole
3	1h 20min to 1h 30min	**Pole cell formation.** Surface cytoplasmic layer becomes thicker and inhomogeneous
4	1h 30min to 2h 30min	**Syncytial blastoderm.** Nuclei divide four or more times. Cortical cytoplasm becomes clearly delimited from the underlaying yolk
5	2h 30min to 3h 15min	**Cell formation.** Cell membranes move down between adjacent nuclei, separating cells
6	3h 15min to 3h 35 min	**Early gastrulation.** Ventral furrow formation along the ventral midline of the embryo
7	3h 35 min to 3h 45min	**Midgut invaginations.** Cephaplic furrow has deepened and is visible from the side
8	3h 45min to 4h 30min	**Germ band extension.** Germ band extends along the dorsal side until the posterior midgut invagination reaches the head region at 65% length
9	4h 30min to 5h 10min	**Stomodeal plate formation.** Cephalic furrow no longer visible.
10	5h 10min to 6h 50min	**Stomodeal invagination.** Anterior midgut anlage moves posteriorly. Ectodermal segmentation becomes apparent as regularly spaced
11	6h 50min to 9h	**Three-layered germ band**. Segmentation is clearly visible. Due to neuroblast proliferation, the dense yolky regions gradually disappear from the head
12	9h to 10h 30 min	**Germ band retraction.** Yolk sac extends to the dorsal side. Anterior and posterior midgut anlagen form visible projections which gradually approach each other
13	10h 30min to 11h 30min	**Shortened embryo.** Germ band completely contracted. Anterior and posterior midgut have fused laterally. The head bends dorsally. Dorsal head ridge formation
14	11h 30min to 13h	**Head involution and dorsal closure.** The germ band stretches anteriorly. Hindgut grows antero-dorsally . Yolk sac is covered by the serosa in the dorsal middle region
15	13h to 15h	**Dorsal closure complete.** Subsequently constrictions divide the midgut into three regularly spaced subdivisions
16	15h until the end of embryonic development	**Condensation of CNS.** Conversion of the sac-like gut into a long convoluted gut. Muscular movements begin in the gut and somatic musculature.

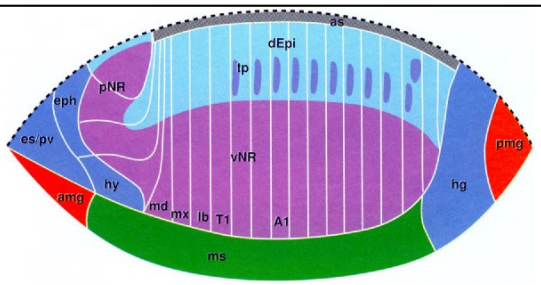

Figure 1.4: Fate map of the *Drosophila melanogaster* blastoderm The fate map is projected onto a planimetric reconstruction of the blastoderm. The upper dashed line represents the dorsal midline and the lower margin represents the ventral midline. A1 Abdominal segment 1; amg anterior midgut rudiment (endoderm); as amnioserosa; dEpi dorsal epidermis; eph epipharynx; es esophagus; hg hindgut; hy hypopharynx; lb labium; md mandible; ms mesoderm; mx maxilla; pmg posterior midgut rudiment (endoderm); pNR procephalic neurogenic region; pv proventriculus; vNR ventral neurogenic region; T1 thoracic segment 1; tp tracheal placodes. Diagram from Hartenstein (1993)

were good to perform cell lineage tracking, as the number of cells of the embryo is relatively small and there are no major morphogenetic movements.

The first method to create a fate map of the *Drosophila* blastoderm were based on the analysis of gynandromorphs (Janning, 1978). Gynandromorphs are genetic mosaics of both male and female cells. Alfred H. Sturtevant analysed many gynandromorphs of *Drosophila simulans* and calculated how frequently two different parts of the embryo were of the same sex. He concluded that two different parts that were more often from different sexes should come from spatially separated cleavage nuclei (Janning, 1978). Importantly, this technique assumes that the position of a cell in the early embryo correlates with its developmental fate. Garcia-Bellido and Merriam improved the work of Sturtevant and using data from 379 gynandromorphs, calculated distances (in "sturt" units, in memory of Sturtevant) between the different embryo parts and construct the first fate map (Garcia-Bellido and Merriam, 1969). Also, histological methods (which consisted in following back to the blastoderm the location of larval organ precursors) and cell ablation methods (killing cells in the blastoderm and correlate its position with the position of the defects detected later) were used to create a fate map of the *Drosophila* blastoderm (Campos-Ortega and Hartenstein, 1985).

In the 1980's José A. Campos-Ortega and Volker Hartenstein combined a labelling techniques (injecting horseradish peroxidase) and histological methods to create a very precise fate map (Campos-Ortega and Hartenstein, 1985), which is still considered a standard modern reference (see Figure 1.4).

Gene expression maps

Techniques such as mRNA in situ hybridization allow to map gene expression patterns directly in the embryo. In situ hybridization is based on labelled probes that are complementary to the mRNA (or DNA) that is wanted to map (Gall and Pardue, 1969). The probe accumulates then only where the mRNA of interest is found. Another technique to map gene expression is the use of a reporter gene. A reporter gene, which codes for a protein that can be easily identified (like the green fluorescence protein or beta-galactosidase), is linked to the regulatory region of the gene of interest so the reporter

gene is going to be expressed where the gene of interest is expressed. Gene expression maps can also be used to create (or refine) fate maps (Gilbert, 2007). For example, if a gene is know to be expressed only in mesoderm precursors, mapping their gene expression in the early embryo will reveal where such mesodermal precursors are located.

Taking advantage of recent high-throughput methods of in situ hybridization (Tomancak et al., 2002; Weiszmann et al., 2009a), the expression pattern of thousands of genes through *Drosophila* embryogenesis have been systematically determined (Tomancak et al., 2002, 2007; Hammonds et al., 2013), and publicly available databases have been developed (Tomancak et al., 2002; Kumar et al., 2011) so any researcher can see where and when a gene is expressed in the embryo.

These databases are suitable for computational image analysis, as the protocols used to produce the images are standardized (Tomancak et al., 2002) and the images can be aligned to an anatomical view (e.g., dorsal, lateral) (Kumar et al., 2011). With the expression patterns of thousands of genes, gene expression maps can be made using clustering techniques, showing regions where the expression of genes is more similar. Frise and collaborators made such analysis, processing thousands of in situ images of the blastoderm embryo and projected them into a virtual representation of the embryo made of ca. 300 triangles (Frise et al., 2010). After clustering the triangles based on their expression similarity, they produce a co-expression map that resembled the fate map shown in Figure 1.4.

Importantly, fate maps (done by lineage tracking, histological or ablation methods) and gene expression maps do not necessarily have to totally coincide. Fate maps inform about which cells in the early embryo will give rise to different cell types or tissues, even when at such early stage the cells can be genetically equivalent.

1.3.3 Gene expression databases of *D. melanogaster*

Berkeley Drosophila Genome Project

The Berkeley Drosophila Genome Project (BDGP) is actually comprised of many projects, whose goals include 1) to complete the high quality sequence of the euchromatic genome of *Drosophila melanogaster* and to generate and maintain biological annotations of this sequence; 2) to produce gene disruptions using P element-mediated mutagenesis; 3) to develop informatics tools that support the experimental process and identify features of DNA sequence; and 4) to characterize the sequence and spatial and temporal expression of cDNAs.

The BDGP insitu project has produced a high-throughput database of mRNA expression in different embryonic stages of *D. melanogaster*, that can be used to complement and extend microarrays or RNAseq analyses (Tomancak et al., 2002). BDGP divides the first 16 stages of embryogenesis into six stage ranges (stages 1-3, 4-6, 7-8, 9-10, 11-12 and 13-16).

A brief description of the hybridization protocol follows, for details see (Tomancak et al., 2002). For the hybridization, they used a set of cDNA clones from the Drosophila Gene Collection (Stapleton et al., 2002), to produce a digoxigenin-labeled antisense RNA probe (Tomancak et al., 2002). Hybridization is carried out in fixed Drosophila embryos in 96-well plates. Successful hybridization plates are mounted on slides to document the expression pattern of each gene with high-resolution digital photographs. Then each image is assigned to one of six developmental stage ranges (Weiszmann et al., 2009b).

Finally, images and annotation data are stored in a modified version of Gene Ontology database. The entire dataset is available to browse or can be download from its webpage (http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl).

In here I used data from the BDGP insitu project whether by directly downloading data from their webpage (e.g., gene expression annotation data) or indirectly from the FlyExpress database (see next section).

Flyexpress

The FlyExpress database (http://www.flyexpress.net/) contains a digitalized library of computationally filtered and standardized images from the high-throughput databases of mRNA expression Fly-FISH and BDGP, and from and peer-reviewed publications. It contains an image-matching search engine that can be used to search for many genes with similar or overlapping patterns of expression in the developing embryo.

The high-throughput databases from which FlyExpress extracts and computationally filter gene expression data differ in the hybridization protocol they use, the number of stages and the staging system, making direct comparisons between them difficult. In contrast with the BDGP database (described in the previous subsection), Fly-FISH uses fluorescence in-situ hybridization probes (Lécuyer et al., 2007) a 17-stage system (compared to a 16-stage system in BDGP) and five stage ranges (compared to six in BDGP). FlyExpress uses a semi-automated pipeline to standardize and align embryos, separating the multi- embryo images of BDGP into single images and discarding partial embryo images (Konikoff et al., 2012). After that, images are assigned to one of three an anatomical views: dorsal, ventral or lateral. Therefore, the expression pattern of a gene in a specific stage and view could be represented in FlyExpress by more than one in-situ image in more than one anatomical view.

In here, from the images available in FlyExpress, I downloaded only those from BDGP, since BDGP uses more stage ranges than FlyFISH and these represent better the whole embryogenesis of *D. melanogaster*. In the Fly-FISH database is focused specially in the early stages, as the last eight developmental stages are contained in one stage range (stages 10-17). I used the FlyExpress database, instead of the BDGP directly, because the standardization protocol used by FlyExpress produces images with embryos in the same orientation and with a cleared background that are more suitable for image computational analysis.

1.4 The Hourglass model in *Drosophila*

In the 19th century, Karl Ernst von Baer stated in his "laws" that within a group of animals the general characteristics appear earlier in development, while the most special appear in late development (see Box 1; von Baer, 1828). This would lead to low morphological variation at early development, gradually increasing as development proceeds.

Other authors (Medawar, 1954; Slack et al., 1993; Duboule, 1994; Raff, 1996) proposed an alternative pattern in which there is great variation in early and late development, while the mid-development would show less variation. This pattern of variation (or conservation) has been called 'phylotypic egg-timer' (Duboule, 1994) and 'developmental hourglass' (Raff, 1996).

Duboule's concept of 'phylotypic egg-timer' was based in the concept of 'phylotypic stage' of Sander (1983), who coined this term to describe the convergence into a conserved segmented germ band stage in insects from very divergent early development (Sander, 1996). In vertebrates, there has been controversy around what should be the phylotypic stage (Ballard, 1981; Slack et al., 1993; Duboule, 1994). Richardson (1995) argued that indeed there is no single conserved stage in vertebrate's development and instead he proposed the term 'phylotypic period' instead.

Initially, two explanations for the hourglass model were proposed. Denis Duboule, after observing that the expression of the Hox genes seemed to coincide with the phylotypic stage, he considered that this could not be a coincidence and proposed that the activation of the Hox genes was the cause for the morphological invariance (Duboule, 1994). In contrast, Rudolf A. Raff proposed that the phylotypic stage was the result of complex interaction between developmental modules at this stage (Raff, 1996).

There is an ongoing discussion about whether the hourglass model (HG), the von Baer law or some other pattern fits the divergence among developmental stages in phylogeny (Richardson et al., 1997; Poe and Wake, 2004; Kalinka and Tomancak, 2012).

Also, it is not clear if the HG, that seems to fit well in vertebrates and arthropods, would apply to other phyla (Raff, 1996) (Salazar-Ciudad, 2010). For example, it is known that the HG model does not apply to spiralians, as many members of this phyla exhibit an early equal cleavage pattern (Henry, 2002).

Recently, the HG have received support from different gene expression studies. Kalinka et al. (2010) used micro-arrays for six Drosophila species and quantified expression divergence at different developmental stages. They found that gene expression was most conserved during the extended germ-band stage (considered the phylotypic period) and that the non-synonymous divergence per site (Dn) correlated with their divergence measures.

They also proposed that the HG pattern is a product of natural selection that acts to conserve patterns of gene expression during mid-embryogenesis (Kalinka et al., 2010). Also, the HG model has been shown to be reflected in the age of the transcriptome. Domazet-Lošo and Tautz (2010) found that when analysing the age of the transcriptome at different stages in the Zebra fish (*Danio rerio*) development (analyzing gene-specific expression data with a phyostatigraphic method), mid-embryonic stages show the older transcriptome (Domazet-Lošo and Tautz, 2010).

Box2. Haeckel, von Baer and the *Naturphilosophie*

In the 19th century, important contributions to embryology were made by advocates of *Naturphilosophie*, a philosophical movement based in Kant and Goethe's ideas, aimed to classify nature into categories or classes. Among their classification efforts, they classified embryological phenomena and draw analogies between embryos of different taxonomic groups (Horder, 2010; Ghiselin, 2005).

The first pattern to be recognized, when comparing developmental trajectories of different species, was the Meckel-Serres law, which proposed that embryos followed a linear succession following the *scala naturae* (a hierarchy of all beings arranged in order of 'perfection', with the man at the top). According to this view, influenced by the *Naturphilosophie*, the embryonic development of a higher organism would be a succession of adult forms of lower organisms (Russell, 1916; Amundson, 2005).

Karl Ernst von Baer K. E. von Baer, a German-Estonian naturalist considered the father of comparative embryology (Russell, 1916), refuted the Meckel-Serres law and formulated his own, known as von Baer's laws (von Baer, 1828). Von Baer's first law state that the more general characteristics of a large animal group (e.g., notochord in chordates) develop before special characteristics (e.g., fur in mammals), while his fourth law state that the embryo of a "higher" animal never resembles the adult of another animal form, but only his embryo.

Importantly, von Baer's views were not evolutionary. The resemblance between developmental trajectories of different species was for him only a reflection of their relationship in the Natural System (Amundson, 2005). Ironically, Darwin used and reinterpreted von Baer's observations on embryonic stages in different species to support common ancestry and therefore, evolution (Darwin, 1859).

Ernst Haeckel Ernst Haeckel supported Darwinism and, in what is known as Haeckel's "Biogenetic Law", said that development (or ontogeny) is a brief summary of the slow and long phylogeny (Haeckel, 1874). In his view, a "higher" organism would pass through a series of conserved developmental stages that represent ancestral forms (this view is known as the "recapitulation theory"). However, in contrast with the Meckel-Serres law, he recognized that this recapitulation was almost never complete, due to evolutionary modifications in development.

> *"The falsification of the original course of development is based to a great extent on a gradually occurring displacement of the phenomena, which has been effected slowly over many millennia, by adapting to the changed conditions of embryonic existence. This displacement can affect both their location and time of appearance. Those former we call heterotopy, the latter heterochrony." (Haeckel, 1903).*

Haeckel's views were more complex than usually acknowledged (Richardson and Keuck, 2002). In fact, he said that it was not that all the mammalian eggs were the same, it was just that with the available tools was impossible to detect the subtle, individual differences, "which are to be found only in the molecular structure" (Haeckel, 1903).

Now is evident that none of von Baer's or Haeckel's hypothesis can be considered "laws", as they are not universal.

In another analysis using Zebra fish, it was shown that the mid-development conservation included that of regulatory regions, with sequences of regulatory regions being most conserved for genes expressed in mid-development; Piasecka et al., 2013).

Studies that have measuring the conservation of genes at the DNA sequence level also seem to support the HG model. Davis et al. (2005) assessed whether proteins expressed at different times during *D. melanogaster* development varied systematically in their rates of evolution (comparing with *D. pseudoobscura*) and found that proteins expressed early in development and particularly during mid-late embryonic development evolve slower. This suggests, according to the authors, that embryonic stages from 12 to 22 hours are highly conserved between *D. melanogaster* and *D. pseudoobscura*, which is consistent with the HG. In a similar study, Mensch et al. (2013) calculated the dN/dS ratio for more than 2,000 genes among six *Drosophila* species, separating genes in three categories: maternal genes (genes whose products are left by the mother in the egg), genes expressed in early development and genes expressed in late development. They found that maternal genes and lately expressed zygotic genes show higher dN/dS ratios (i.e., are less conserved) than early expressed zygotic genes. Finally, it has also been found that genes expressed in the adult have higher dN/dS ratios than genes expressed in the pupa and those of the pupa have higher dN/dS ratios that those expressed in the embryo (Artieri et al., 2009).

Some limitations of these last studies is that they classify the genes in a few broad temporal categories that do not permit to precisely determine the temporal dynamics of conservation and that are based only in divergence data (dN/dS ratios between two species). A study that integrates polymorphism data from natural populations would improve the evolutionary interpretation of these patterns, as it would allow to estimate what proportion of the dN are adaptive (as explained in section 1.2.2). Measuring adaptation is specially relevant as some authors have argued that the HG is caused by different selection pressures in early and late development (Slack et al., 1993; Kalinka and Tomancak, 2012; Wray, 2000)

1.5 Ciona

The ascidian *Ciona intestinalis*, a marine invertebrate animal, has a long history in developmental biology and evoutionary biology. Darwin highlighted the importance of the ascidians due to their close phylogenetic relationship to the vertebrates (REF). Also, it provided one of the first evidences of localized determinants of cell specification (Conklin, 1905). Although their adult form is a sessile filter feeder, its tadpole larva has characteristic features of the chordate group: a dorsal neural tube, a notochord surrounded by muscle and a ventral endodermal strand (Satoh, 1994). Ascidians show morphogenetic movements during gastrulation and neurulation similar to vertebrates and both share common genetic regulators of cell specification (REF). Their relative short life cycle, almost transparent body and rapid development facilitate many genetic techniques and are partly responsible for the re-emergence of *C. intestinalis* as model organism in developmental biology (Levin et al., 2012).

1.5.1 Current knowledge about Ciona development

The sequencing of the *C. intestinalis* genome (Dehal et al., 2002) facilitated its comparison with other vertebrate sequenced genomes and the analysis of gene expression through its life cycle. The *C. intestinalis* genome is only 160Mb and contains 16,000 genes, a gene number similar to the invertebrate *D. melanogaster* genome and only is half of the genes found in some vertebrates (REF). This low number of genes (compared to vertebrates) can be explained by the finding that many gene families or subfamilies have only one representative in *C. intestinalis* (Dehal et al., 2002). Relevant efforts have been made to describe the spatial expression patterns of individual genes (REF). The spatial expression patterns of >1,000 cDNA clones have been described using whole-mount in situ hybridization techniques at different developmental stages (Imai et al., 2004). Importantly, the developmental stages included cover a wide temporal range, e.g., blastula, gastrula and tapole stages (REF). Taking advantage of the ascidian invariant cleavage pattern and well described lineage analysis (Conklin, 1905; Nishida, 1987), the spatial expression of many genes have been described at the single cell level up to the early gastrula stage (REF), making this an invaluable resource to investigate the spatio-temporal dynamics of gene expression.

1.5.2 *Ciona intestinalis* life cycle

Developmental stages

Quick intro of developmental stages at 18°C, based on (Hotta et al., 2007).

1.5.3 The ANISEED database

The ANISEED database (2015 version) integrates expression data from large-scale in situ hybridization studies with embryo anatomical data of ascidians (Tassy et al., 2010; Brozovic et al., 2016). This database includes 27,707 *Ciona intestinalis* gene expression profiles by in situ hybridisation for approximately 4500 genes acquired from more than 200 manually curated articles (Brozovic et al., 2016).

The expression data is represented using an ontology-based anatomic description of the embryos. The ANISEED database also includes expression data from the Ghost database in *Ciona intestinalis* (Satou et al., 2005), which contains the spatial expression patterns of >1,000 cDNA clones by whole-mount in situ hybridization at different developmental stages. Taking advantage of the invariant ascidian cleavage pattern and well-described lineage analysis (Conklin, 1905; Nishida, 1987) the cDNA spatial expression have been described at the single cell level up to the early gastrula stage (Imai et al., 2004).

The ANISEED database also includes biometry data (e.g., volume, surface/volume) and 3D embryo models (at a single-cell resolution) of ascidian embryos until the early gastrula (Tassy et al., 2006). Importantly, combining the gene expression and 3D embyo models at a single cell resolution it is possible to reconstruct the gene expression pattern in 3D, as I did here. As the ANISEED database do not contains 3D embryo models for tailbud stages, in order to do reconstruct the gene expression in 3D in these stages, I used a 3D model of *C. intestinalis* mid tailbud anatomy at a single cell resolution (Nakamura et al., 2012)(see methods and study II).

Table 2. Embryonic stages of *C. intestinalis* and main morphological characteristics (based on Hotta et al., 2007)

Stage	Description	Time after fertilization	Morphological characteristics
1	**Zygote**	0min to 55min	From the fertilisation event up to the end of the first mitotic cycle
2-5	**Early cleavage**	55min to 3h	Five mitotic divisions, until the 32-cell stage. First and second cleavages separate the left and right halves, and the anterior and posterior halves, respectively.
6-9	**Late cleavage**	3h to 4.5h	Very small B7.6 cell pair in the posterior end. Asymmetric divisions in the vegetal hemisphere. Embryo flattens on its vegetal side. Vegetal cells take a columnar shape
10-13	**Gastrula**	1.5h to 6.3h	Invagination and migration of endodermal and mesodermal cells inside the embryo. At the early gastrula, the vegetal side of the embryo takes a horseshoe shape. Embryo starts elongating anteriorly
14-16	**Neurula**	6.3h to 8.5h	The embryo, with an oval shape, continues elongating. Notochord precursors intercalation and convergence. Neural tube formation and closure (starting from the posterior side)
17-20	**Initial and early taibud**	8.5h to 10h	Clear separation between trunk and tail. Neuropore closure. Tail starts bending
21-22	**Mid tailbud**	10h to 11.9h	Intercalation of the notochord cells is completed. The tail bends ventrally so the embryo adopts a half-circle shape. Length of the tail twice as long as the trunk
23-25	**Late tailbud**	11.9h to 17.5h	Pigmentation of the otolith can be observed. Palps formation. Vacuolization of notochord cells. The tail is bent dorsally
26	**Hatching**	17.5h	The larva hatches. Head adopts an elongated rectangular shape

Aims of the study

In this work, I have analysed publicly accessible spatio-temporal gene expression data of two model organisms, *Drosophila melanogaster* and *Ciona intestinalis*, together with population genomics data of *D. melanogaster*. Using a statistical approach, I address three questions:

i. How do complexity and compartmentalization increase in the embryo during development?

ii. Can adaptation be found in specific anatomical parts of the embryo or developmental stages?

iii. Is the Hourglass model (which states that there is less amount of inter-specific variation in mid-development; see section 1.4) supported by evidence of natural selection at the DNA sequence level?

These questions have been selected for the great interest they have aroused in the scientific community since the early days of developmental and evolutionary biology. The work presented here is based on and uses concepts from three main biology fields: developmental biology, evolutionary biology and population genetics. Nowadays, the combination of these scientific fields constitute multiple research programmes. Modern evolutionary developmental biology (evo-devo) is the explicit combination of the first two fields.

Material and Methods

3.1 On the complexity measures used in this work

In the following paragraphs I will describe the three measures of complexity I used in here: relative area/volume, disparity and roughness. I consider them complexity measures because they inform about specific features in embryonic development that are intuitively associated to the increasing complexity of the organism. The relative area/volume of expression relates to the notion of the progressive compartmentalization of the embryo, disparity informs about how the different parts of the embryo become more different (at the genetic level) from each other, and the roughness measure relates to the notion of genes being expressed in progressively more complex spatial expression patterns (as defined by the shape of the gene expression pattern).

Compartmentalization

As mentioned before, the process of embryo compartmentalization, or the subdivision of the embryo in different parts during development, is considered to be largely reflected in the expression of the genes in progressively smaller areas. Therefore, a good measure of compartmentalization would be to measure the relative area of expression of all genes during development. Gene expression patterns are usually visualized as a 2D image, which reflects the distribution (seen after an experiment such as mRNA in situ hybridization) of a gene product from a specific anatomical view (e.g., lateral, dorsal). More recently, methods like 3D imaging technique Optical Projection Tomography (Sharpe, 2003; Summerhurst et al., 2008) allow to record gene expression patterns in 3D.

In the case of 2D images, the measure of compartmentalization I use in here is the relative area of the expression of a gene divided by the area of the whole embryo (pixels with expression divided by all the pixels of the embryo image). Therefore, I called this measure "relative area" which ranges from 0 (no expression) to 1 (ubiquitous expression). In the case of a gene expression in 3D, the compartmentalization measure becomes the relative volume of expression, which also ranges from 0 to 1 and is calculated as the volume of the cells/tissues with expression divided by the whole embryo volume. Of course it could be said that the compartmentalization of an embryo should not only be reflected in the decrease of the relative area of expression of the genes during development, but also should be reflected in a greater proportion of genes expressed differentially in different parts of the embryo. To account for this, I used a "disparity" measure (explained below), that would be informative in the mean degree of dissimilarity between the different parts of the embryo.

Disparity

The disparity measure is related to McShea's "number of part types" complexity measure. Ideally, the number of cell types, if these could be precisely known, would be a good measure of complexity during development. However, as mentioned before, there is no clear criteria to determine when (at the genetic level) a new cell type is formed during development. Usually, genetic markers are used to determine if a cell (or its

descendants) can be considered a specific cell type. This could be the case of anterior or posterior compartment cells within *D. melanogaster* para-segments. For example, the use of wingless (*wg*) or sonic hedgehog (*hh*) in situ hybridization can be used to define if a cell is part of the anterior or posterior compartment. It could be said, however, that the earliest presence of expression of these genes marks the *fate* of the cell, more than the actual cell type. In fact, before the anterior and posterior segment compartments are completely defined, the expression level of these genes changes over time and form a positive feedback to stabilize each other expression. Therefore, a cell that early expresses *wg* would not necessarily differentiate into a posterior compartment if its neighbour para-segment does not express *hh* (Bejsovec and Martinez Arias, 1991).

Instead of trying to determine the number of cell types during development, an alternative approach would be to quantify how different at the gene expression level are the cells in the embryo. In *D. melanogaster*, as it was not possible to have expression for each individual cell, I divided the 2D embryo in "regions" of approximately the same area using polar coordinates (Fig. 3.1A; see study I). In the case of *C. intestinalis*, I used expression data at a single cell level for the early stages and at tissue level for the tailbud stages (see study II).

Then, I decided to use pearson's correlation as similarity measure between cells/tissues or regions. With this method, the expression of all the genes available in two cells can be compared, giving a value of 1 if all genes have exactly the same expression (their expression is completely correlated), to -1 if all the genes show opposite gene expression (their expression is anti-correlated). An advantage of this method is the possibility to include the information of all available genes, diminishing a possible bias in the gene selection. A further application of this method is the use of clustering algorithms (e.g., hierarchical clustering) to find the main temporal patterns of gene expression in microarray analysis (Eisen et al., 1998; Wen et al., 1998). Also, this method can be used to separate cell types based on the genetic differences between them (Gohlmann and Talloen, 2009).

The pearson's correlation matrix is a measure of gene expression similarity between genes. In here, I was interested on quantifying the difference (i.e., disparity) in gene expression between cells, not on their similarity. Therefore, the disparity between two cells becomes: disparity = 1 - pearson's similarity. The disparity measure ranges from 0 to 2, with a 0 value when the gene expression between cells is exactly the same.

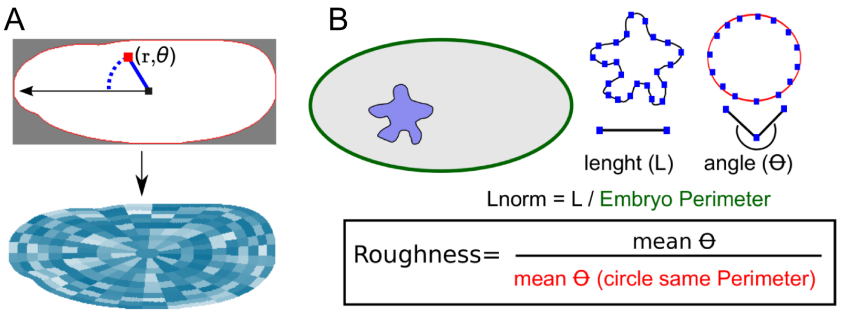

Figure 3.1: Polar regions and 2D roughness measure. A) The embryo of each stage was divided in 257 regions using polar coordinates. The embryo for stage 11-12 is shown with the 257 regions in a random color. B) A schematic embryo (gray) with a gene expression pattern in blue. Roughness is the mean major angle (θ) between each node (at every L length pixels in the contour) and its two immediate neighbours, normalized by the mean angle of a circle of the same perimeter.

Roughness

The roughness measure analyses the complexity of the shape of a gene expression pattern. In the case of 2D images, the shape of the gene expression pattern can be extracted as a closed outline formed by the boundaries of gene expression, while in the case of 3D representations, the shape of the expression pattern would be the 3D external surface of the union of the cells that are expressing such gene.

Therefore, a gene expression pattern reflects necessarily the spatial distribution of the cells expressing such gene. When analysing and comparing the shape of diverse gene expression patterns, i.e., the cells/tissues with expression are different between genes, there is an obvious impossibility to determine landmark points (whether around a 2D outline or 3D surface) that would establish a clear one-to-one correspondence between them. This could be done in the case of comparing the expression pattern of a single gene at a specific developmental stage between different individuals. Therefore, a landmark-free method (like outline methods for 2D or surface methods for 3D data) is best suited to deal with the type of data analysed in here.

There are practically no studies in the literature that have quantified and compared the shape of gene expression patterns in a systematic manner (one exception is the recent study of Martínez-Abadías et al., 2016). In here, I will consider that a gene expression pattern is complex if based on the curvature of its 2D contour or 3D surface. For 2D gene expression patterns, I used a "roughness measure", that is similar to the shape function $\phi^*(l)$ used in eigenshape analyses (Lohmann, 1983), as it measures how much the curvature of a closed outline deviates from the angles of a circle of the same perimeter (Fig. 3.1B; see study I). To calculate the roughness of a expression pattern I first selected points in the contour every L (length) pixels. Then, vectors between each node and the two immediate neighbour nodes in the contour are calculated and the biggest angle formed between them is measured. The roughness value is then the mean angle normalized by the mean angle of a circle of the same perimeter.

I selected the roughness measure instead of some other measure outline based method like Fourier analysis because the roughness value gives an intuitive descriptor of complexity, i.e., a value of 1 would be a simple "circle-like" shape, and a value greater than 1 would mean a higher curvature of the outline. McLellan and Endler (1998) compared various measures of spatial complexity applied to the outlines of the leaves of many tree species. They included a "margin roughness" measure that is very similar to the one I use here (the difference is that they does not normalize by the mean angle of the circle nor he uses different lengths of vectors) and found that there was no marked differences between the margin roughness and a Fourier analysis with up to 64 harmonics (both performed equally well). On the contrary, one advantage of the Fourier analysis would be that it is an information preserving algorithm (Pavlidis, 1980) i.e., it is possible to reconstruct the shape after the analysis, a feature that was not considered relevant for this study. Other feature of the roughness measure is that allows to measure the complexity of shape at different spatial scales (something similar can be done by varying the number of harmonics used in Fourier analysis; McLellan and Endler, 1998). Therefore, it can be tested not only if the complexity of the gene expression shape increases during development, also if this increase is the same at different spatial scales.

In order to use a similar measure of curvature in 3D, I used the Dirichlet normal energy (DNE; described briefly in section 1.1.3 which quantifies the deviation of a surface from being planar (Fig. 3.2; see study II). Importantly, both measures are normalized

to remove size and orientation effects.

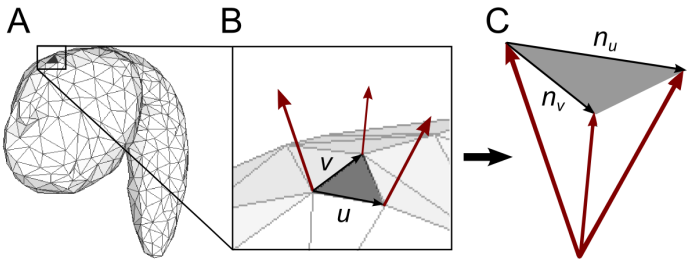

Figure 3.2: Dirichlet Normal Energy (DNE). A) A surface mesh representing a a mid-tailbud embryo in C. intestinalis. B) DNE calculates the energy value $e(p)$ of each polygon (like the one in grey) in the surface. The polygon is characterized by vectors u and v, which represent the edges of the polygon. Then, normal unit vectors are estimated as the normalized average of normal vectors of the triangle faces adjacent to each vertex (red arrows). C) If vertex normals are translated to a common origin point, their end points form a polygon with edge vectors nu and nv, which represent the spreading of nu and nv. In a simplistic way, DNE can be defined as the spreading of nu and nv relative to the spreading of u and v (Bunn et al., 2011; Winchester, 2016).

It is important to mention that the aim of this analysis is not to discern which mechanisms (e.g., cell-cell signalling or morphogenetic movements) are responsible for the changes in complexity of the shape of gene expression pattern, but to quantify how this happens during embryonic development.

The relationship between these measures

In this work I use three different measures of complexity: relative area/volume, disparity and roughness (analysed with DNE for 3D expression patterns), which are informative of different and independent aspects of complexity. For example, a decrease in the area/volume of gene expression should not be necessarily correlated with an increase in disparity, as the genes that are reducing their expression area could be restricted to the same part of the developing embryo. Only in the case of an embryo with all genes showing ubiquitous expression, there is a clear relationship between disparity and relative area/volume, as the relative area of expression and disparity would be 0 and 1 respectively. If there are however, many genes expressed in only a part of the embryo, these measures are not necessarily correlated.

This can be illustrated with a simple example shown in Fig. 3.3, in which there are different alternative gene expression scenarios of an imaginary embryo with six cells. In each scenario, the embryo expresses four genes in different relative areas (i.e., in a different number of cells). The mean relative area is 0.5 for all scenarios, but the mean disparity varies in a two-fold manner. In the scenario that shows the largest disparity, each cell expresses a unique combination of genes, while the scenario with the lowest disparity, 4 of 6 cells do not have a unique expression profile.

The roughness and disparity independence can be easily exemplified in the case of a blastula. Blastula is the name to define the multicellular aggregate stage that results from the subdivision (cleavage) of the zygote. The blastula shape topology and geometry is usually simple (Forgacs and Newman, 2005), typically consisting of a ball of cells with an interior cavity (called "blastocoel"). If in a spheric blastula, composed

of also spheric cells (like that of a sea urchin) a large proportion of genes would be expressed ubiquitously and a small proportion of genes would be expressed in single cells, both roughness and disparity would be relatively low. However in the case of a large proportion of genes expressed in different single cells, and a low proportion of genes expressed ubiquitously, the disparity would be high, but the roughness would be very similar than in the previous case. This would be because the roughness quantifies the shape of the expression pattern, irrespective to size. Therefore the roughness of a gene expression in a single spheric cell would be practically the same that the roughness of a gene expression in the whole spheric embryo.

The independence of the roughness measure with the size of the gene expression (i.e., relative area/expression) comes from the roughness normalization. The normalization of the 2D roughness consists of dividing the mean angle of a gene expression pattern by the mean angle of a circle with the same perimeter, and in the case of 3D roughness (i.e., DNE) it consists on transforming the 3D expression surface into a polygonal surface mesh with a determined number of polygons.

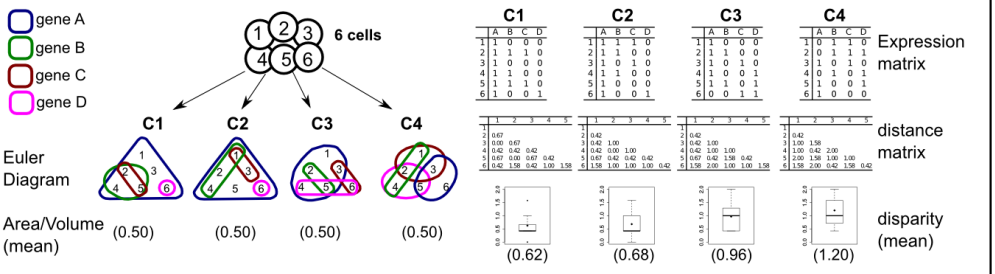

Figure 3.3: Relation between area/volume of expression and disparity measures. An embryo of six cells (top left) is shown expressing four different gene expression combinations (C1, C2, C3 and C4) of four genes (A, B, C and D). All combinations have a mean relative area/volume of 0.5. Each gene expression configuration is represented as an Euler diagram (representing the subset of the cells in which it is expressed in a color code shown at the top left) and as binary expression matrix (top right). The pairwise distance between the cells, calculated as 1-(pearson's correlation), is shown as a matrix. At the bottom right, a distribution plot of the pairwise distances of each combination and the mean disparity are shown below in parenthesis.

3.2 Data mining and handling

The work presented here is based on the analysis of publicly available data contained in many databases, shortly introduced in the Review of the Literature section. In the next paragraphs I will describe how the data used in here was acquired and processed, for the details, see the corresponding study.

In situ Hybridization data

***D. melanogaster* (study I and III)** Images were downloaded from the FlyExpress Database version 5.1 (Kumar et al., 2011) on February 2013. Only genes with laterally oriented images for the six stages used in BDGP (Tomancak et al., 2002) were considered.

Images were systematically retrieved and resized to 320 x 128 pixels (as in Konikoff et al., 2012) with ad-hoc Perl scripts. The gene expression pattern was obtained using an

adaptive threshold based on the mean and variance of a grey-scale version of each image. Three different threshold were used, resulting in three different filtered gene expression patterns of each original in situ image. I visually selected the image that resembled the most to the original gene expression pattern. Genes with ubiquitous expression in stages 1-3 and 4-6 were considered as entirely black images.

To correct for small variation in the shape of the embryos I adjusted each embryo to an stage ideal embryo shape, with an algorithm I made to morphometrically deform the real embryo contour of each image to the corresponding stage ideal shape. Finally I applied a "smoothing filter" to produce a smooth expression pattern and eliminate isolated white/black pixels. I also manually filtered images from the literature or directly from BDGP of Transcription Factors or Growth Factor genes that did not have information in FlyExpress. These manually filtered images were also morphometrically deformed with the algorithm mentioned above. The resulting dataset contained 1218 genes with expression information in the six stages used in BDGP.

In addition to the whole-mount in situ RNA-hybridization images, the BDGP database contains, for each gene, the list of the embryonic anatomical structures in which such gene is expressed (Tomancak et al., 2007). Each gene expression is described by one or several of those of anatomical terms by an expert. This information was retrieved from the BDGP downloads page (http://insitu.fruitfly.org/insitu/html/downloads.html/), which contains the annotations of almost 8,000 genes. We removed genes with only "no staining" as anatomical term, leaving a total of 5762 genes.

***C. intestinalis* (study II)** I downloaded the in situ hybridization data (ish.zip file) from the download section of the ANISEED database on 28th of December 2015. The expression data for the first three stages is at the cell level, while in the tailbud stages is at the tissues or specific regions of the embryo level. I extracted the information of the 32 cells, 64 cells, 112 cells, early tailbud, mid tailbud and late tailbud stages. Only expression data from experiments reported to have Wild type phenotype, "public" publication status, with in situ hybridization as experiment design and whose probe was assigned to a Kyoto Hoya (KH) (Satou et al., 2008) gene model. I excluded data from experiments whose image characterization was reported as "not sure" or too broadly as "part of whole embryo".

The number of genes analyzed is n=745 for the 32-cell stage, n=758 for the 64-cell stage, n=809 for the 112-cell stage, n=1082 for the early tailbud, 1092 for the mid tailbud and 887 for the late tailbud.

3.2.1 Transcriptomics ans population genomic data

modENCODE (study III and IV) Gene expression levels in reads per kilobase per million mapped reads (RPKM) units for 30 developmental stages were retrieved from Gelbart et al. (2013), who analyzed RNA-seq throughput data from the modENCODE project (Graveley et al., 2011).

For using RNA-seq data to compare expression between samples, a normalization step was performed to adjust for varying sequencing depths and other potential technical effects across replicates (see study III)

DGRP (study II and IV) The population genomic data comes from 168 inbred lines of *D. melanogaster* sequenced in the Freeze 1.0 of the Drosophila Genetic Reference Panel

(DGRP) project (Mackay et al., 2012). The DGRP population was created collecting gravid females from a single population of Raleigh, North Carolina (USA), and following the full-sibling inbreeding approach during 20 generations to obtain full homozygous individuals. DGRP lines showing high values of residual heterozygosity (>9%) that were observed to be associated with large polymorphic inversions (Huang et al., 2014) were not included.

Due to requirements of the software used to estimate the rate of adaptive substitution, the original data of 168 lines set was reduced to 128 isogenic lines by randomly sampling the polymorphisms at each site without replacement. Finally, residual heterozygous sites and sites with no quality value were excluded from the analysis.

3.3 Estimating adaptation

CHECK BIBLIOGRAPHY AND TEXT IN GENERAL OF THE WHLE SECTION!

To estimate adaptation during *D. melanogaster* embryogenesis, I used the DFE-alpha method (see section 1.2.2; Eyre-Walker and Keightley, 2009), which infers adaptation combining polymorphism and divergence data. In the last years, different population-genomic projects have sequenced, in different species, the genome of many individuals of a population (or a set of populations) (The 1000 Genomes Project Consortium, 2010; Mackay et al., 2012; Pool et al., 2012; Wallberg et al., 2014), providing a valuable resource of genomic polymorphism data at the population level.

3.3.1 Population genomic data

In this study, I used data from the *Drosophila melanogaster* Genetic Reference Panel (DGRP) (see section 1.2.3; Mackay et al., 2012), which consists of inbred *D. melanogaster* lines. Importantly, the lines are derived from a single outbred population, so they capture natural variation (as genetic polymorphism) and are ideal to use with methods like the DFE-alpha.

More specifically, the DGRP data comes from 168 inbred lines of a North American population of *D. melanogaster*(DGRP Freeze 1, Mackay et al. 2012).

The program used for estimating the rate of adaptation (DFE-alpha, Eyre-Walker and Keightley 2009 ;see below) needs all sites to have been sampled in the same number of chromosomes, so the original dataset was reduced to 128 isogenic lines by randomly sampling the polymorphisms at each site without replacement.

We used the release 5 of the Berkeley Drosophila Genome Project (BDGP 5, http://www.fruitfly.o as the reference genome (REF). The divergence statistics were estimated from a multiple genomic alignment between DGRP lines and *D. yakuba* using BDGP 5 coordinates (publicly available in http://popdrowser.uab.cat REF).

The number of sites and substitutions and the folded site frequency spectrum (SFS) were computed using an ad hoc Perl script.

3.3.2 Estimation of natural selection

To estimate the rate of adaptation we used the DFE-alpha method and the software provided with it. The program is intended to estimate several population genomic

parameters (alpha, etc) in a group of genes rather than in a single one, because estimates in an individual gene are noisy and can be affected by the lack of segregating (divergent) sites. Because of that, for performing the analysis, we randomly sampled groups of genes with replacement to estimate the rate of adaptive evolution (see Bootstrapping, Methods). As the neutral reference, we used the positions 8-30 of short introns (less or equal than 65 bp) following (Heyn REF), to determine the rate of adaptation in a particular sequence (in this case, 0-fold degenerate sites). For a validation analysis, we also used 4-fold degenerate sites as a proxy for the neutral mutation rate. The MKT does not take into account the segregation of slightly deleterious alleles, biasing downward the adaptive rate, while the DFE-alpha method does and, thus, provides a more accurate estimation than the MKT and other methods that do not take polymorphism data into account (REF Eyre-walker) .

The DFE-alpha method, in contrast to the dN/dS ratio, allows the estimation of the proportion of non-synonymous substitutions that are actually adaptive, neutral or slightly deleterious. Finally, to estimate the rate of adaptive evolution using DFE-alpha, it is necessary to combine data from several genes because estimates from a single gene are noisy and often undefined because of the lack of segregating (or divergent) sites for some site classes.

Results and Discussion

4.1 Comparative study between *Drosophila* and *Ciona* (I and II)

4.1.1 Compartmentalization

I estimated the degree of compartmentalization calculating the relative area or volume of expression of genes during development. My intention here was not to focus on individual genes, but to get a global overview of the embryo compartmentalization and differentiation processes based on expression data of thousands of genes, i.e., using a statistical approach.

One would expect, and it has been implicitly assumed (Carroll et al., 2001) (Davidson, 2001) that the compartmentalization of the embryo (as I measure it here) increases during development. However, the specific temporal dynamics of this increase in any species is not known. Neither is clear if the dynamics should be similar for different species, or for different groups of genes. As the development of *Ciona* and *Drosophila* are very different and it would be impossible to compare them stage-by-stage, I focused here in three major developmental periods: pre-gastrula, gastrula, and post-gastrula stages. These periods are easily recognizable in both species facilitating the comparative analysis.

I found that in both species, the relative area or volume decreased in a non-linear way (see Figs X). However, the timing of the major decrease was different. In *Drosophila* the major decrease occurred at very early development, from maternal to early gastrula stage (Fig. 4.1). Practically half of the genes in follows this decrease pattern: 46% of the genes were characterized as having a non-linear decrease in their relative area. In contrast, in *Ciona* the volume of expression decreases mostly after gastrulation (between the 112-cell and the early tailbud stage). However less dramatic, I found significant differences between the 32-cell and 64-cell stages, and between the 64-cell and 112-cell stages.

The difference in the timing of the major change on compartmentalization between species must relate to differences in their specific development. The earlier compartmentalization of *Drosophila* is most probably due to its derived early development, namely,

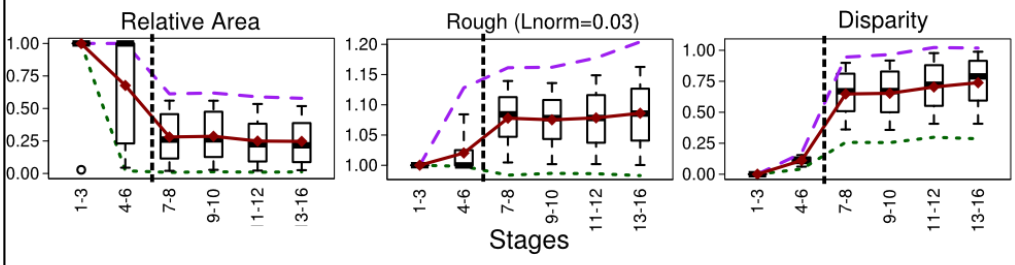

Figure 4.1: Measures in *Drosophila*. Distribution plot of the relative area of expression (left), roughness (center) and disparity (right) for all genes in each stage. Diamonds represent the mean, boxes the Inter Quartile Range (IQR). Whiskers 10 and 90 percentiles. Dashed line represents the max values and dotted line the min values (mean of the last and first decile, respectively). Stages on the x-axis, vertical dashed line represents gastrulation entry.

the syncytial blastoderm. During the blastoderm stage, approximately 4,000 cell nuclei can 'communicate' with each other only by TFs (Jaeger, 2011). The direct cross regulation of gene expression facilitates a rapid and highly dynamic process which seems to be responsible for the early spatial restriction of a great proportion of developmental genes. In contrast, *Ciona*'s early embryonic patterning is based on maternal determinants and signalling events mostly between neighbouring cells (Lemaire, 2009), which act in a combinatorial way (Hudson et al., 2007) to establish a unique TF combination in more than half of the blastomere pairs before gastrulation (Imai et al., 2006) determining most of their fates. Thus, even when in *Ciona* most of the cell fates are already determined (by the specific combination of a fraction of TFs) and the embryo can be said to be already highly compartmentalized, this is not evident at the global level of gene expression, which I am measuring here.

Therefore, the 'delay' of compartmentalization observed in *Ciona* could be explained by the relatively slower process of signal transduction (as in *Ciona*) compared to the gap gene network (in *Drosophila*).

4.1.2 Disparity

As the relative area (or volume) of expression informs on how genes are expressed in progressively smaller regions in the embryo, the disparity can inform about how different regions of the embryo express increasingly different combinations of genes. Therefore, both measures reflect slightly different aspects of complexity that are independent from each other. A decrease in the volume of expression of genes does not necessarily imply an increase in spatial disparity: genes could decrease their volume of expression but end up restricted to the same parts of the embryo. If the majority of genes would be expressed ubiquitously (this is large volume), however, then the mean disparity between its regions would be necessarily low.

My results show that in each species, the global disparity pattern is similar to the relative area or volume patterns. Therefore, in *Drosophila* the disparity increases mostly in the transition from the maternal to early gastrula and in Ciona this major change occurs after gastrulation.

It is important to notice that these measures should not necessarily correlate, as it could be that between two stages the relative area of expression decreases but not the

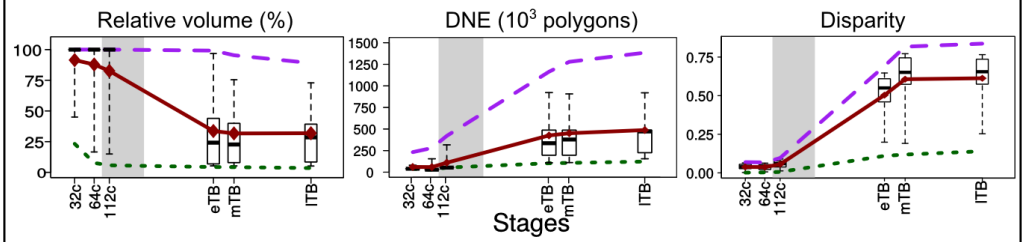

Figure 4.2: Measures in *Ciona*. Distribution plot of the relative volume of expression (left), DNE (center) and disparity (right) for all genes in each stage. Diamonds represent the mean, boxes the IQR. Whiskers 10 and 90 percentiles. Dashed line represents the max values and dotted line the min values (mean of the last and first decile, respectively). Stages on the X-axis (s32c, 32-cells; s64c, 64-cells; s112c, 112-cells; eTB, early tailbud; mTB, mid tailbud; lTB, late tailbud). Grey area represents gastrulation period.

disparity if the genes are expressed in the same part of the embryo or vice-versa. In *Ciona* I found an example of such case, when there is no perfect correspondence between the relative volume and the disparity of expression: disparity increased significantly between early to mid-tailbud stages but no significant differences between the relative volume of expression of these stages were found (II, Fig. 3A). This means that, on average, genes are expressed in a similar number of tissues in these stages, but in the mid tailbud the combination of genes expressed in these tissues are more different between each other.

This shows that the disparity measure is useful specially when is complemented with the relative area (or volume) measure to describe the compartmentalization of the embryo.

4.1.3 The leading role of TFs and GFs (and other signalling molecules)

I wanted the test in both species if TFs and GFs showed an earlier compartmentalization or greater disparity when compared to the rest of the genes. This would be expected from their allegedly leading role in early pattern formation.

Using a GOterm analysis in *Drosophila*, I found that TFs (GO:0003700) and GFs (GO:0008083) showed smaller relative area of expression that the rest of genes in the blastoderm stage (Fig. 4.1). The TFs are also expressed in smaller areas than the rest of the genes in all subsequent stages, while the GFs are expressed in smaller areas at the blastoderm (stage 4-6) and extended germ band stages (stage 9-10 and 11-12) (I, Fig.4). A similar result for TFs was reported in Drosophila by Hammonds et al. (2013). They made an extensive analysis of TFs expression using manual annotation of gene expression based on an anatomical controlled vocabulary and classifying every gene as ubiquitous, patterned, ubiquitous-patterned, or maternal (from the BDGP database; Tomancak et al., 2007). They found that the fraction of TFs expressed in a restricted pattern (assigned to a tissue) was higher, when compared to other genes, in all zygotic stages with the exception of the stage 13-16.

The results I show for stages 4-6, 7-8, 9-10 and 11-12 are consistent with Hammonds et al., as the higher proportion of the TF genes showing a restricted or tissue-specific expression pattern would imply that TFs are expressed in smaller areas in the embryo. For the 13-16 stage, contrary to these authors, I showed that the TFs are highly compartmentalized. This might indicate a limitation of the annotation method used by Hammonds et al., to capture the high spatial compartmentalization of the TFs in this stage.

In *Ciona*, I performed a similar analysis using the categorization of TFs and signaling molecules (SIGs) made by Imai et al. (2004). SIGs consist of genes of receptor tyrosine kinase (RTK) pathways such as FGFs and intracellular signalling molecules such as MAPK, Notch, Wnt, TGFβ, Hedgehog and genes in the JAK/STAT pathways (Imai et al., 2004).

As expected, TFs volume of expression decreased faster than non-TFs. The TFs showed lower volume of expression in the 64-cell and 112-cell stages (II, Fig. 3B). The results are similar for maternal and zygotic genes (maternal/zygotic classification based on Matsuoka et al., 2013; II,Fig. S1). I then compared TF families and found that six TF families showed lower relative volume in the early gastrula (BZIP, T-box, bHLH, HMG, Nuclear Receptor, and 'Other-TFs') but only T-box genes showed a lower relative volume from the 32-cell stage until gastrula (II,Fig. S2).

The results obtained for the T-box gene family (conserved in metazoan and several non-metazoan lineages (Sebé-Pedrós et al., 2013)) are consistent with the known important role these genes have in diverse metazoan species early cell fate specification (reviewed in: Papaioannou, 2014; Showell et al., 2004. Examples of T-box genes in *Ciona* are Tbx6 and *brachyury*, crucial for muscle tissue formation (Mitani et al., 1999; Nishida, 2005) and for notochord specification (Yasuo and Satoh, 1998), respectively. I also found that the SIGs showed significant lower relative volume of expression than the rest of the genes in the 32-cell, 64-cell, and 112-cell stages (II, Fig. 3B). Specifically, in the 64-cell stage RTK-MAPK, Wnt and TGFβ families showed significant higher disparity in the 64 cells stage, suggesting a predominant role of these pathways in the patterning of the embryo at this stage. This is consistent with known short range induction events by nodal and various FGFs, which are part of the TGFβ and RTK-MAPK signalling pathways, respectively (Lemaire et al., 2008).

In general, the fact that in these two species that display a very different development TFs and GFs (or SIGs in the case of *Ciona*) are more compartmentalized than the rest of the genes precisely in the stage before entering gastrulation, is consistent with these genes having a special role in pattern formation and compartmentalization. Therefore, my results support the hypothesis of the leading role of TFs and GFs in driving pattern formation and compartmentalization in the early embryo.

4.1.4 2D and 3D roughness analyses

I wanted to test the hypothesis of gene expression spatial patterns becoming more complex during development. In here, with gene expression spatial pattern I mean the spatial distribution of the cells or tissues expressing a specific gene.

Considering that I had information in 2D in *Drosophila* and in 3D in *Ciona*, it was necessary to apply a specific method for each species. For *Drosophila*, I developed a 'roughness' measure (Salvador-Martínez and Salazar-Ciudad, 2015) which accounts for the curvature of the contour in a 2D gene expression pattern, normalizing it with the contour of a circle of the same perimeter. In *Ciona*, I applied a similar measure of curvature in 3D, called 'Dirichlet normal energy' (DNE), which quantifies the deviation of a surface from being planar (Bunn et al., 2011). Both measures not only inform about the overall imbrication or convolution of the shape of a gene expression pattern, but also do it at different spatial scales.

In the following paragraphs, to improve the readability of the text, I will refer to the roughness measure implemented in *Drosophila* as 2D roughness and to the DNE measured used in *Ciona* as 3D roughness.

The results show that both 2D and 3D roughness increase in a non-linear way during development. As with the compartmentalization and disparity, what changes between species is where the major change is found.

In *Drosophila*, the major change is found in the transition from the blastoderm to the early gastrula (Fig. 4.1). When analysing the maximal values (mean of the last decile) it can be seen that they increase initially in the pre-gastrula, reach a stationary phase at mid-embryogenesis and finally increase in the last stages. As I mentioned in the literature review (section X), the maximal values are informative about the overall morphological spatial complexity of the embryo in a given stage. When comparing roughness at different spatial scales (I, Methods), I found that in the last three stages

the roughness values are significantly higher at smaller spatial scales is significantly higher that at the higher spatial scales.(Fig. S2 in article I).

In Ciona, the 3d roughness increase throughout development (II, Fig. 5), with the major change between the 112-cell and the early tailbud (Fig. 4.2) The max (mean of the last decile) values increase substantially already between the 64 and 112 cells stages (with 1000 and 10000 polygonal faces), while the min values (mean of the first decile) remain practically constant during development, showing that the most complex patterns in each stage get increase their DNE value but there is always a proportion of very simple expression pattern. Also, I found that at low spatial scales (1000 and 10000 polygons per mesh; II, Fig. 5) I found that the mean DNE of the late tailbud is higher than at the mid tailbud (one-way ANOVA pvals < 0.05).

In summary, this results show that the complexity of distribution in space of cells/tissues expressing a gene increases through development, and that these complexity (measured with the 2D and 3D roughness) increase in both *Ciona* and *Drosophila* in a similar way than the other two measures, compartmentalization and disparity. Also, by analysing the roughness at different scales, I found compelling evidence that complexity may be increasing not only through all the development but also that it does at finer spatial scales over time.

4.1.5 Synexpression territories

I wanted to explore in both species the relative degree of similarities between different parts of the embryo within and between different developmental stages.

To do this I used two different approaches, based on the differences between the databases I used for each species. In *Drosophila* I used the polar regions with which I computationally divided the embryo. In *Ciona*, I took advantage of the available information at the individual cell/tissue.

In both cases I used a clustering algorithm to produce dendrogram representing the relative degrees of similarities between all regions of different stages at the same time (Fig. 4.3 and Fig. 4.4). I will refer to the regions that clustered together as 'synexpression territories' (STs).

In *Drosophila*, after cutting the dendrogram at a specific threshold and filtering STs with less than 50 genes expressed with a minimum specificity (see methods in I for a detailed description), 30 STs were selected for further analyses (Fig. 4.3 B).

Finally, I grouped the STs in eight 'meta-territories', as I wanted not only to see how the regions in the embryo formed different STs, but also how different STs cluster with each other, as this is informative of the degree of differentiation between stages. If STs cluster with other STs in the same stage, it would mean that the majority of genes change their expression in a similar way over time independently of where they are. If STs cluster with other STs in the same part of the embryo in successive stages, it would mean that this part of the embryo has expression dynamics independent from other parts of the embryo, which would be expected in already differentiated cells/tissues.

The results show that stages 1-3 and 4-6 each one form a ST. If a cut-off is selected so that stage 4-6 is divided in four sub-territories (I,Fig. S3) the embryo splits in four parts: anterior, posterior, dorsal and ventral. This correspond to a nearly Cartesian system one could expect from the two signalling systems known in the earliest patterning in *Drosophila* (the A/V and D/V signalling cascades; (Gilbert, 2014)). The STs seem to coincide with the known embryo fate map (see Fig. 4.3 D; Hartenstein, 1993) and many

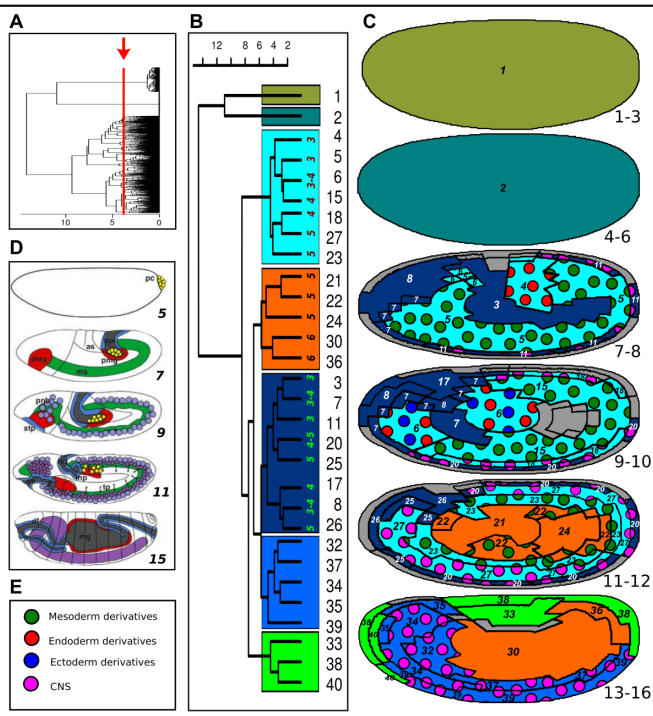

Figure 4.3: Synexpression territories (ST). (A) Dendrogram produced by hierarchical clustering on a similarity matrix (pearson's correlation) of all the embryo regions of the six stages. Red line shows the cut-off to produce 40 STs. (B) Dendrogram reconstructed using only territories with at least 50 genes with a minimum specificity (I, methods). The coloured boxes show the main branches of the dendrogram. The number indicated inside the boxes represent the stages each ST corresponds to (3 is stage 7-8, 4 is stage 9-10, 5 is stage 11-12 and 6 is stage 13-16). The ST number is at the right. (C) STs mapped onto the embryo. Gray regions have less than 50 genes expressed. Background color refers to which 'meta-territory' (in B) each ST is part of. Coloured circles represent GOterm enrichment of a specific tissue/germ layer derivative (shown in E). Stages in the lower-left part of each embryo. From stage 7-8, the ST number (as in B) is indicated. (D) Hartenstein's embryo schemes (Hartenstein, 1993) with their respective stages in the left upper part. (E) Colour code of specific tissue/germ layer derivative used in C.)

of them are enriched with GOterms that coincide with their expected fate. For example, in stage 7-8 (just after gastrulation) there is a ST that corresponds spatially with the germband and is enriched with mesodermal GOterms (Fig. Fig. 4.3 C).

There are two meta-territories that appear in the last stage (light blue and green, Fig. 4.3 C), which suggests that the tissues/organs related to those STs differentiate quite late. One meta-territory is enriched with terms related to epidermis such as cuticle development ('chitin catabolic process' [GO:0006032] and 'cuticle development' [GO:0042335] STs 33 and 38), which coincides with cuticle deposition by epithelial cells during stage 16 (Ostrowski et al., 2002). The other meta-territory corresponds spatially with the CNS of the embryo and is indeed enriched with CNS GO-terms. The CNS territory is enriched with GOterms like 'dendrite morphogenesis' (GO:0048813) and 'axon guidance' (GO:0007411).

In *Ciona*, because gene expression information in tailbud stages is based on tissues

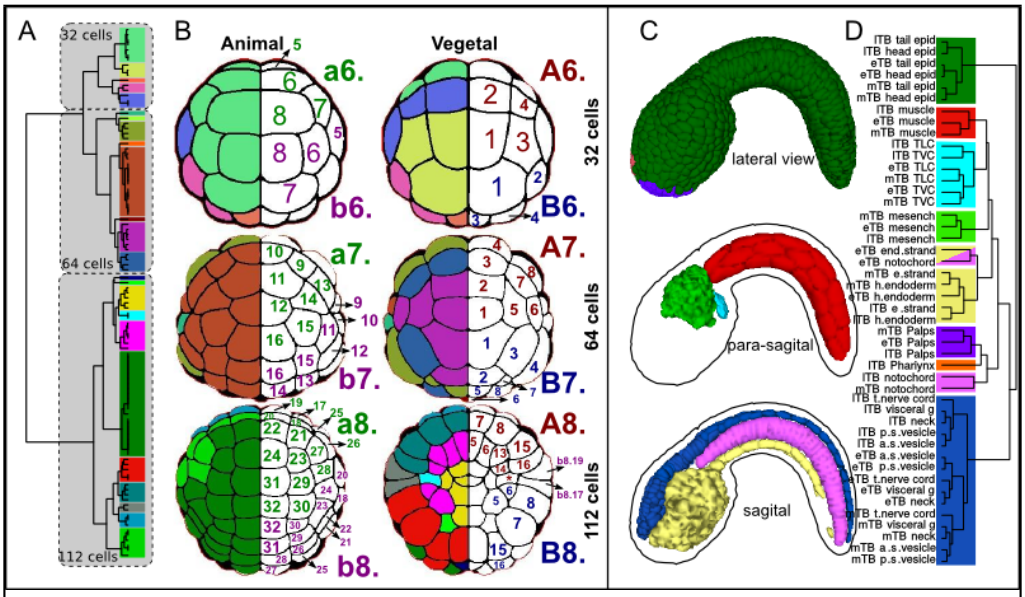

Figure 4.4: Ciona synexpression territories. (A) Dendrogram produced by hierarchical clustering of cells in 32-cell, 64-cell and 112-cell stages. Dashed boxes show that STs cluster by stage. Coloured boxes show the cut-off to produce 24 STs. (B) Names of cells (Conklin nomenclature; Conklin, 1905) indicated with a prefix shown at right. STs in the 32 cells, 64 cells and 112 cells stages (top, middle and bottom, respectively). Colour refers to which ST of the dendrogram (in A) each cell is part of. Animal view based on Nicol and Meinertzhagen (1988) and vegetal view based on Cole and Meinertzhagen (2004). The cell marked with a star (*) is the A7.6 cell, that in this analysis represents their descendant cells (A8.11 and A8.12). C) Dendrogram produced by hierarchical clustering of tissues in early, mid and late tailbud stages. The coloured boxes show the cutoff to produce 10 STs. (D) STs in the tailbud stages shown in a lateral, para-sagital and sagital views of a mid tailbud 3D embryo model (from Nakamura et al., 2012). Colour refers to which ST of the dendrogram (in C) each tissue is part of.

and not on individual cells as the early stages, I analysed the STs of these stages separately (II, Methods).

If in the early stages, three 'meta-territories' are formed, each one would correspond to one stage, i.e., STs in early stages cluster by stage. Thus, even if at the first three stages a high proportion of blastomeres express a nearly unique combination of transcriptional factors (Imai et al., 2006), the bulk change in gene expression is common to all blastomeres. Within each early stage, STs coincides very well with the know fate map (II, Fig 6A; II, Fig. S8), with some exceptions I will describe in the next subsection.

In contrast, in tailbud stages practically all STs cluster by tissue/cell type, which indicates that the in early tailbud, most tissues are already quite differentiated. This is consistent with studies analysing these stages at the level of individual or small sets of genes (Corbo et al., 1997; Di Gregorio and Levine, 1999).

My analysis in the early stages is similar to the one made by Imai et al. (2006), who used the expression profile of 53 zygotically TFs in single cells in the 16, 32, 64, and 112-cell stages, to perform a hierarchical clustering (for each stage separately). It is different in two aspects: I performed the clustering using the blastomeres of different stages and my analysis is not restricted to TFs. As I said previously, using various stages is informative of the overall differentiation process and can be used to discern between

differentiation scenarios, as the differences between early and tailbud stages I found here.

The main difference between species is that, in *Drosophila* the differentiation process continues throughout whole embryogenesis (as new STs were formed until the last stage I analysed) and different organs differentiate at different developmental times. In contrast, the *Ciona* embryo seems to be already genetically differentiated at the early tailbud (as the STs of all the tissues in the tailbud stages cluster together) so the last embryo stages consist only of moderate morphogenetic movements (mainly cell elongation; Hotta et al., 2007). Therefore, the ST analysis is a valuable tool, based on differential gene expression, to get a global perspective on the local differentiation of the embryo.

4.2 Main spatio-temporal profiles of gene expression in *Drosophila* (I)

With a time series cluster analysis (Ernst and Bar-Joseph, 2006) of the relative area of expression, I found the eight main spatio-temporal profiles of gene expression in the embryonic development of *Drosophila* (I, Fig. 5). As expected, the most common profile (n=297 genes) follows the global profile of non-linear decrease in the first stages (I, Fig 5).

Among the rest of profiles, I found both linear increase and decrease profiles and a 'hill-like' profile (initial increase and further decrease with the higher values at stage 7-8) The linear decrease profile (n=167 genes) was enriched with 'mitotic cell cycle' (GO:0000278), 'RNA processing' (GO:0006396) and 'chromatin modification' (GO:0016568) GOterm genes, highlighting biological processes that first are present in the whole embryo and become more and more restricted in space as development proceeds. The 'mitotic cell cycle' term, for example, most likely relates to the fast mitotic cycles in the earliest embryo. During stage 1-3 nine fast and synchronic mitotic divisions take place in the entire embryo, then in stage 4-6 mitotic divisions 10-13 occur more slowly, almost synchronically. The 14th cycle, zygotically controlled, is long and of different durations in the embryo.

With a temporal co-expression cluster analysis using microarray data through the life cycle of *D. melanogaster*, Arbeitman et al. (2002) found that most cell cycle genes were expressed at high levels during the first 12h, but only a few are expressed at high level thereafter. My analysis is consistent with this, as I found that the profile of linear decrease (I, Fig. 5A) is enriched with such genes. In this sense, this study is complementary to Arbeitman et al., and adds the spatial dimension to their temporal expression profiles.

4.3 Discrepancies between fate map and STs (II)

I found a few cases in *Ciona* in which cells with the same fate where contained in different STs. This would be the case of: 1) cells whose fate is disproportionally affected or determined by a small number of genes (as this analysis reflect quantitative differences at the level of hundreds of expressed genes but can not distinguish between the relative importance of each gene) or 2) cells that although having a restricted fate at a certain stage their differentiation is not complete (at the level of gene expression).

This analysis could not be made in *Drosophila* as the gene expression data is not a the single level resolution.

An example of the latter is a ST in the 112-cell stage (in magenta; 4.4 B; II, Fig. S8) that contains precursors of the notochord (A8.5, A8.6, A8.13, and A8.14, B8.6) and mesenchyme (B8.5) (Tokuoka et al., 2004). The latter come from a secondary notochord/mesenchyme bipotential cell (B7.3). It has been reported that the expression of Twist-like 1, necessary for mesenchyme differentiation, starts at this stage (Imai, 2003). This evidence, together with the inclusion of the mesenchyme cell in this otherwise exclusively notochord territory (primary and secondary), seems to indicate that the differentiation of cell pair B8.5 as mesenchyme is still incomplete at this stage.

Gene expression dynamics in cell-lineages

I analysed the gene expression similarity between lineage-related cells (i.e., between daughters cells and between mother/descendants cells) in the early stages (II, Fig. 8). In general, cells are more closely genetically to their sister cells than to their mother/descendants, which is reflected in the clustering of STs by stages discussed before. I found also that at the 64-cell stage, cells that show more genes expressed differently than their ancestors are neural fated cells, which might be related with the fact the unrestricted state of these cells at this stage (i.e., their descendants will give rise to different cell fates).

4.4 Adaptation in *Drosophila* embryogenesis (III and IV)

I combined the Synexpression Territories (STs) approach with genome-wide coding-region polymorphism data (from the DGRP database) and the coding-region divergence between *D. yakuba* and *D. melanogaster* in order to estimate the proportion of adaptive non-synonymous substitutions (ω_α) in the genes expressed in each ST (n=589 genes; III, Methods).

Using this approach, I could chart a spatial map of natural selection acting on *Drosophila*'s embryo anatomy. I complemented this with a analysis using available annotation of gene expression (n=2,835 genes) using a controlled vocabulary of anatomical structures from the BDGP database (Tomancak et al., 2007).

The results showed a few STs with significant higher or lower ω_α (permutation test; III, Methods)

4.4.1 STs or anatomical terms with high ω_α

STs 13 and 32 (ST number comes from the hierarchical clustering algorithm; see Fig. 4.5), which showed a higher ω_α, seem to correspond to the forming foregut and hindgut (stage 11-12) and to the CNS (stage 13-16) respectively. To explore if ST 32 high ω_α was indeed related to the CNS, I separated the genes CNS or not-CNS related. I found that both groups showed a high ω_α, which suggests that in addition to the CNS, another structure in the anterior region would be under positive selection. Using the anatomical terms approach, no anatomical terms related to the CNS were found to have high ω_α with the initial criteria. I therefore applied a more stringent criterion to

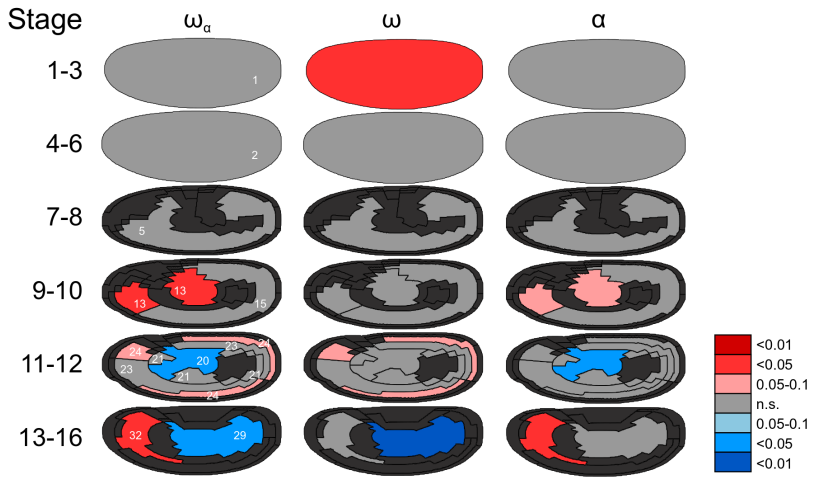

Figure 4.5: ω_α on embryonic territories over space and time. Territories drawn in red in the central column mark significantly high ω_α while those in blue mark significantly low ω_α in space in each of the 6 developmental stages (rows). Other columns depict α, the proportion of base substitutions fixed by natural selection, and ω, the rate non-synonymous substitutions relative to the mutation rate. Territories in dark gray are territories without enough specific genes to be analyzed. The statistical was calculated by a permutation test using all the genes analyzed (see Material and methods). Territory 13 in stage 9-10 (ω_α: 0.059, p = 0.045). Territory 20 from stage 11-12 (ω_α: 0.022, p = 0.048; α: 0.259, p = 0.028). Territory 24 from stage 11-10 (ω_α: 0.070, p = 0.061). Territory 29 from stage 13-16 (ω_α: 0.037, p = 0.047; ω: 0.074, p < 0.001). Territory 32 from stage 13-16 (ω_α: 0.068, p = 0.044; α: 0.71, p = 0.04).

consider genes as part of an anatomical term (before a gene could have a maximum of seven anatomical terms associated instead of a more stringent number of three) and found that 'Embryonic brain' showed high ω_α (permutation test, p = 0.046). Also, with the anatomical terms approach, I found that genes associated with 'Gonads', in the last stage, clearly showed evidence of adaptive evolution (III, Figure 2), which is consistent with previously reported high rates of adaptive substitution in the testes (Akashi, 1994; Civetta and Singh, 1995; Nuzhdin et al., 2004; Pröschel et al., 2006)

Finally, ST 24 (stage 11-12; Fig. 4.5) that seems to corresponds to part of the trunk mesoderm marginally significant high ω_α (p = 0.061). A similar result was obtained with for the anatomical term 'Trunk mesoderm' in stage 9-10 (p = 0.087).

4.4.2 STs or anatomical terms with low ω_α

STs 20 and 29 with showed low ω_α (Fig. 4.5) seem to correspond to the forming midgut (stage 11-12) and to the forming larval digestive system (stage 13-16) respectively. Similar results are found when using the anatomical term approach, as low ω_α was found in many anatomical terms related to the digestive system in the last stage: 'Embryonic midgut', 'Embryonic salivary gland', 'Embryonic hindgut', 'Embryonic proventriculus'. Also, combining three related anatomical terms, 'Embryonic foregut', 'Embryonic epipharynx' and 'Embryonic hypopharynx', that separately did not have enough genes to be considered in the analysis, showed low ω_α.

The lack of adaptive change in the forming digestive system might reflect their rela-

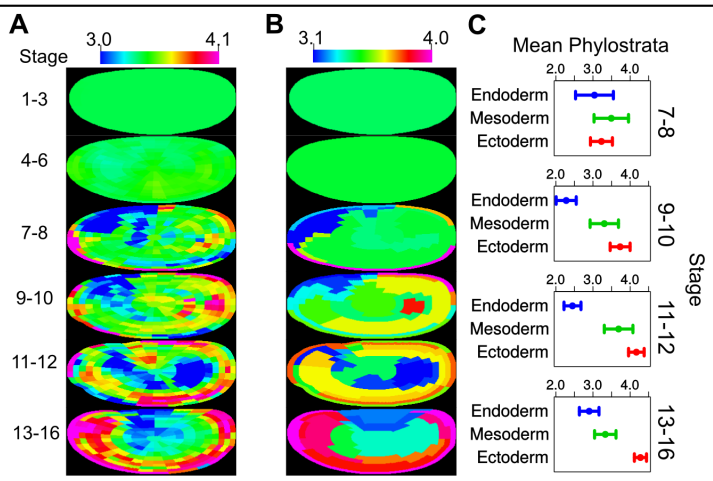

Figure 4.6: The center of the embryo expresses older genes. (A) Heatmaps showing the transcriptome age index (TAI) im polar regions (B) Heatmaps showing the TAI for STs. (C) Mean phylostrata of genes assigned to each germ layer. Circles represent the mean and whiskers the SEM.

tive enrichment in metabolic genes (Marianes and Spradling, 2013). The coding regions of metabolic genes have been found to be more conserved than non-metabolic genes (Peregrín-Alvarez et al., 2009).

4.4.3 Transcriptome age index and other genomic determinants

Then, I wanted to test if the 'age' of the genes, also differed between different parts of the embryo and how this related to the adaptation results. For that, I used the phylostratigraphic maps of *D. melanogaster* (Drost et al., 2015), that assign a phylogenetic age to each protein-coding gene based on the phylogenetic level at which orthologs for a gene are found (a young found only in Drosophilids would be very young, of age 1). Also, I used a modified version of the Transcriptome Age Index (TAI; Domazet-Lošo and Tautz, 2010) and applied it to the polar regions and STs (see III, methods; Fig. 4.6). TAI is low for regions expressing old genes and large for regions expressing young genes. I found that STs with low ω_α express (on average) older genes (high TAI values; see Fig. 4.6). Similar results were found for anatomical structures (III, Fig. S2). I also found that in stage 13-16 the mean phylogenetic age of the genes expressed in the endoderm is lower than in other germ-layers, specially compared to the ectoderm (Fig. 4.6). Similar TAI results between germ-layers were found by Domazet-Loso et al. (2007) but without stages comparisons. The correlation between adaptation and gene age fit the expectation that older genes would more likely perform essential functions than younger genes, and that, as older genes would have been moulded by natural selection for longer times would be therefore more close to optimality (assuming that this function is conserved). Therefore, more room for changes would be expectable in embryo regions with a larger proportion of younger genes.

I also found that embryo polar regions with high ω_α have low codon bias (previously reported by Sharp, 1991; Betancourt and Presgraves, 2002; Haerty et al., 2007) and that, as Plotkin and Kudla (2011) previously found, regions high codon bias show have high

levels of gene expression (average RNAseq levels per region; III, methods). To clarify the relation between these three variables, I fitted a multivariate linear regression and found that embryo regions with high ω_α exhibit low codon bias relative to what would be expected from their gene expression levels (III, Fig 5). The negative correlation between codon bias and protein adaptation that I found would be expected given that, an adaptive aminoacid change in a protein would be probably different from a change that would increase codon usage efficiency (Hershberg and Petrov, 2008; Presnyak et al., 2015).

4.4.4 Selective constraint in late embryogenesis

Analysing the RNAseq developmental data from the modENCODE project (Graveley et al., 2011) with the DFE-alpha method (IV, methods), I found that from hour 10 until 24 of embryogenesis show significant low ω_α and ω (see Fig X in section X), which would be consistent with the low rate of adaptive change seen in many anatomical structures in stage 13-16 (as stage 13-16 of BDGP roughly maps to RNA-seq samples em10-12 hr, em12-14 hr, em14-16 hr, and em16-18 hr of modENCODE; Hammonds et al., 2013). Therefore, by combining different approaches, I could identify that the proteins produced in late embryogenesis change less their aminoacid sequence (i.e., are more conserved). This phenomenon, of some proteins evolving slower, has been called 'selective constraint' and has been linked to the higher degree of functionality of such proteins (Kimura, 1983). Most importantly, I could identify which specific anatomical structures expressed genes with a higher degree of conservation.

4.5 Adaptation trough *Drosophila* life cycle (IV)

Taking advantage of the modENCODE developmental data (Graveley et al., 2011), which contains expression data for 30 stages of the whole life cycle of *D. melanogaster* (12 embryonic at 2-h intervals for 24 h, 6 larval, 6 pupal and 3 sexed adult stages at 1, 5 and 30 days after eclosion), ω_α and other evolutionary rates for the genes expressed in each stage (for details see: IV, methods) were estimated. For each stage, ω_α, ω, ω_d and α (Fig. 4.7) were estimated with the differentially expressed genes (genes with expression different from zero and excluding those expressed in all stages).

Also, different 'genomic determinants' (codon usage bias, intron length, or number of exons) were estimated for each stage, in order to test how their temporal patterns would relate to the temporal patterns of the estimated evolutionary rates (i.e., ω_α, and others).

The results show a clear temporal pattern, in which ω_α, ω_d and ω show their highest value at the first embryo stage (0-2hr) to then decrease until the 10-12hr embryo stage. Their values remain low through most of embryonic development (the embryonic period were briefly discussed in section 4.4.4). Then, from larval stage L3 the values increase suddenly and remain relatively high through all the pupa. Finally, in the adult stages, males show similar values to those of the pupa, but females show significantly lower values ($p < 0.001$).

α follows a similar temporal profile to that of ω_α, but the differences in α values through the life cycle are smaller than those of ω_α.

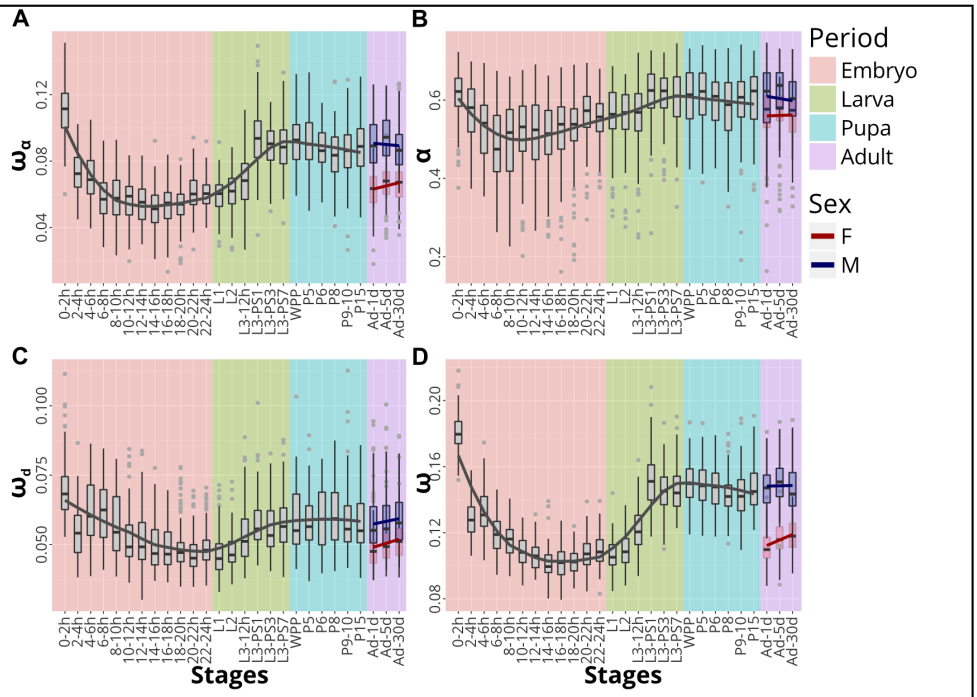

Figure 4.7: ω_α **(A),** α **(B),** ω_d **(C),** ω **(D) through the life cycle of** ***D. melanogaster*** Each time point represents 1,000 random samples of 350 genes (with replacement) expressed in a stage. Red line represents a LOESS regression. Female and male stages are fitted to a linear regression. There are 12 embryonic stages at 2hr intervals (from 0h to 24h). Larval stages at first instar (L1), second instar (L2) and third instar (L3). L3 stages are subdivided into the first 12 hours (L3-12h) and several puff stage (L3-PS1 to L3-PS7). WPP is the white pre-pupae stage. Pupal stages 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 (Ad-1d, Ad-5d, Ad-30d).

Similar results were found when considering genes with a two-fold difference between maximal and minimal expression (IV, Fig S1), after excluding immune system and testes genes (IV, Fig. S2) or when the mutation rate is estimated using the 4-fold degenerate sites (IV, Fig. S3).

Most genomic determinants follow a temporal profile that is either very similar to that of ω_α or the opposite of it (IV, Figs 2 and 3).

Messenger complexity (number of transcripts divided by the number of exons) follows a very similar pattern to ω_α (rank correlation $r^2 = 0.622, p = 2.54x10^{-6}$ with the male stages and $r^2 = 0.7, p = 1.67x10^{-6}$ with female stages). The temporal pattern of gene size, number of exons, codon usage bias and number of transcripts per gene patterns is the opposite of that ω_α for males (rank correlations: $r^2 = 0.702, p = 1.66x10^{-6}$; $r^2 = 0.867, p = 6.19x10^{-7}$; $r^2 = 0.781, p = 1.27x10^{-6}$; $r^2 = 0.667, p = 1.85x10^{-6}$, respectively) and females (rank correlations: $r^2 = 0.697, p = 1.68x10^{-6}$; $r^2 = 0.888, p = 4.67x10^{-7}$; $r^2 = 0.813, p = 1.04x10^{-6}$; $r^2 = 0.756, p = 1.44x10^{-6}$, respectively).

Using a fuzzy clustering algorithm, I found that genes that are expressed at high levels in the earliest development that rapidly decrease their expression to very low

levels (cluster 1 and 2; IV, Fig. 4) are likely responsible for the high ω_α, ω_d, ω and *alpha* values in the first embryonic stages (IV, Fig. 4). Also, I found that a subset of genes whose expression increases only in the last stages of embryonic development (cluster 8; IV, Fig. 4) showed highω_α and ω(permutation test, p = 0.008).

Cluster 1 and 8, which showed high ω_α also showed significantly low gene size, number of exons and number of transcripts (permutation test, $p < 0.001$).

We find that all pupa and adult male stages exhibit the highest levels of adaptive change while mid and late embryonic stages show high conservation.

Concluding Remarks

The study of organismal complexity during embryonic development presented here shows that there are commonalities and differences between *D. melanogaster* and *C. intestinalis*. Both species showed a non-linear increase in all complexity measures, while the most remarkable difference is the timing of the major change in complexity, which is earlier in *D. melanogaster* (around gastrulation). Another common pattern is the early increase in complexity when considering only transcription factors or growth factors (or other signalling molecules). This confirms the special role these genes have in early metazoan development. It could be therefore expected that the evidence presented here, regarding these type of genes, should be observed also in other species.

One important result of this work is that within each species, the three complexity measures showed a similar pattern (even when it would not be necessarily the case; see section X). This means that altogether, these measures (compartmentalization, disparity and roughness) are reflecting a global pattern of increase in complexity in each species. Therefore, it could be hypothesized that a similar increase in complexity would be found using alternative measures of complexity (e.g., spatial entropy). Further analysis would be required to test this hypothesis. Also, the Synexpression Territories analysis allowed to "reconstruct" the main embryonic differentiation events in both species in a consistent manner with the current knowledge of the development of these model organisms and without focusing in specific genes.

The elaboration of an adaptation map on the fruit fly embryo can be considered a proof of concept of how the combination of diverse fields like evolutionary developmental biology and population genomics, and new techniques such as the phylostratigraphy, can be useful to give a fresh view on an old problem. Using these maps, it was possible to visually identify that the center (internal part) of the embryo expresses a more conserved and older transcriptome, while the outside (external part) expresses phylogeneticaly younger and less conserved genes. This evidence seems to support the hypothesis of the antecedence of the endoderm with respect to the ectoderm (Hashimshony et al., 2014). It would be interesting to extend this adaptation mapping analysis for the entire development (until the adult stage) as it could be that in later stages, different structures or organs have been under positive or negative selection.

The estimation of adaptation over the entire life cycle of *D. melanogaster*, as presented here, supports the HG model of development. We find, as other analyses previously have, that the mid-embryogenesis is highly conserved. The work presented here is different from previous ones in that it uses a more complete spatio-temporal dataset and a method that uses inter and intra-specific DNA coding variation to estimate, with an unprecedented precision, the proportion of adaptive changes.

Furthermore, as a result of this work is hypothesized that the hourglass model can be partially explained by various genomic features. However, further work is necessary to test this hypothesis.

OPEN QUESTIONS AND FUTURE DIRECTIONS

The emergence of new techniques, like "spatial transcriptomics" of tissue sections at single-cell resolution (Stahl et al., 2016) could make possible to have information, derived from a single experiment, of all the genes expressed in a 2D section of an embryo. The application of the measures presented here could be apply to this kind of data in a straightforward manner, solving the limitation in resolution of the work presented here.

It is important to mention that this work has used differential gene expression in the embryo and its spatial distribution as a tool to investigate complexity. However, embryonic development can not be reduced to differential gene expression. Cellular behaviours and the physical properties of the cells and tissues have also a causal role in the developmental process. It would be interesting to be able to measure the differential apportionment to complexity increase of the different developmental mechanisms.

The increase of organismal complexity and the study of adaptation during development remain fascinating topics after many centuries, and still offer many open questions to be solved. The incredibly fast pace of data generation, the development of new techniques and sophisticated methods give hope to finally open the black box of development.

References

Abzhanov, A. et al (2006). The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature 442*(7102), 563–7.

Adami, C. (2002). What is complexity? *BioEssays 24*(12), 1085–1094.

Adami, C. (2004). Information theory in molecular biology. *Physics of Life Reviews 1*(1), 3–22.

Adami, C., Ofria, C. and Collier, T.C. (2000). Evolution of biological complexity. *Proceedings of the National Academy of Sciences 97*(9), 4463–4468.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics 136*(3), 927–35.

Alberts, B. et al (1994). *Molecular Biology of the Cell* (3rd ed.). Garland Science.

Amundson, R. (2005). *The Changing Role of the Embryo in Evolutionary Thought: Roots of Evo-Devo*. Cambridge Studies in Philosophy and Biology. Cambridge University Press.

Anfinsen, C.B. (1973). Principles that Govern the Folding of Protein Chains. *Science 181*(4096), 223–230.

Apter, M. and Wolpert, L. (1965). Cybernetics and development I. Information theory. *Journal of Theoretical Biology 8*(2), 244–257.

Arbeitman, M.N. et al (2002). Gene expression during the life cycle of Drosophila melanogaster. *Science (New York, N.Y.) 297*(5590), 2270–5.

Arias, A.M. (2008). Drosophila melanogaster and the Development of Biology in the 20th Century. In C. Dahmann (Ed.), *Drosophila: Methods and Protocols*, pp. 1–25. Totowa, NJ: Humana Press.

Arthur, W. (2010). *Evolution: A Developmental Approach*. Wiley.

Artieri, C.G., Haerty, W. and Singh, R.S. (2009). Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of Drosophila. *BMC biology 7*(1), 42.

Ballard, W.W. (1981). Morphogenetic Movements and Fate Maps of Vertebrates. *American Zoologist 21*(2), 391–399.

Barton, N.H., Briggs, D.E., Eisen, J.A., Goldstein, D.B. and Patel, N.H. (2007). *Evolution*. Cold Spring Harbor Laboratory Press.

Batty, M. (1974). Spatial Entropy. *Geographical Analysis 6*(1), 1–31.

Batty, M. (2010). Space, scale, and scaling in entropy maximizing. *Geographical Analysis 42*(4), 395–421.

Batty, M., Morphet, R., Masucci, P. and Stanilov, K. (2014). Entropy, complexity, and spatial information. *Journal of Geographical Systems 16*(4), 363–385.

Bejsovec, A. and Martinez Arias, A. (1991). Roles of wingless in patterning the larval epidermis of Drosophila. *Development (Cambridge, England) 113*(2), 471–85.

Bell, G. and Mooers, A. (1997). Size and complexity among multicellular organisms. *Biological Journal of the Linnean Society 60*(3), 345–363.

Betancourt, A.J. and Presgraves, D.C. (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), 13616–20.

Bonner, J.T. (2004). Perspective: The Size-Complexity Rule. *Evolution 58*(9), 1883–1890.

Bookstein, F.L. (1997). Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis 1*(3), 225–243.

Brozovic, M. et al (2016). ANISEED 2015: a digital framework for the comparative developmental biology of ascidians. *Nucleic Acids Research 44*(D1), D808–D818.

Bunn, J.M. et al (2011). Comparing Dirichlet normal surface energy of tooth crowns, a new technique of molar shape quantification for dietary inference, with previous methods in isolation and in combination. *American journal of physical anthropology 145*(2), 247–61.

Campos-Ortega, J.A. and Hartenstein, V. (1985). *The Embryonic Development of Drosophila melanogaster*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Canning, E.U. and Okamura, B. (2003). Biodiversity and Evolution of the Myxozoa. *Advances in Parasitology 56*, 43–131.

Carroll, S.B., Grenier, J.K. and Weatherbee, S.D. (2001). *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. Malden, MA.: Blackwell Science.

Casari, G., Sander, C. and Valencia, A. (1995). A method to predict functional residues in proteins. *Nature structural biology 2*(2), 171–8.

Chaitin, G.J. (1999). *The Unknowable*. Discrete Mathematics and Theoretical Computer Science. Springer Singapore.

Christiansen, J.H. et al (2006). EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic acids research 34*(Database issue), D637–41.

Civetta, A. and Singh, R.S. (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), 1085–95.

Cole, A.G. and Meinertzhagen, I.A. (2004). The central nervous system of the ascidian larva: mitotic history of cells forming the neural tube in late embryonic Ciona intestinalis. *Developmen-*

tal biology 271(2), 239–62.

Conklin, E.G. (1905). The organization and cell-lineage of the ascidian egg. *Journal of the Academy of natural sciences of Philadelphia 13*, 1–119.

Corbo, J., Erives, A., Di Gregorio, A., Chang, A. and Levine, M. (1997). Dorsoventral patterning of the vertebrate neural tube is conserved in a protochordate. *Development 124*(12), 2335–2344.

Crombach, A., Cicin-Sain, D., Wotton, K.R. and Jaeger, J. (2012). Medium-throughput processing of whole mount in situ hybridisation experiments into gene expression domains. *PloS one 7*(9), e46658.

Dailey, L., Ambrosetti, D., Mansukhani, A. and Basilico, C. (2005). Mechanisms underlying differential responses to FGF signaling. *Cytokine & Growth Factor Reviews 16*(2), 233–247.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* John Murray.

Davidson, E.H. (2001). *Genomic Regulatory Systems: In Development and Evolution.* Academic Press.

Davidson, E.H. (2009). Developmental biology at the systems level. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1789*(4), 248–249.

Davis, J.C., Brandman, O. and Petrov, D.A. (2005). Protein evolution in the context of Drosophila development. *Journal of molecular evolution 60*(6), 774–85.

Dehal, P. et al (2002). The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science (New York, N.Y.) 298*(5601), 2157–67.

Di Gregorio, A. and Levine, M. (1999). Regulation of Ci-tropomyosin-like, a Brachyury target gene in the ascidian, Ciona intestinalis. *Development 126*(24), 5599–5609.

Domazet-Loso, T., Brajković, J. and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in genetics : TIG 23*(11), 533–9.

Domazet-Lošo, T. and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature 468*(7325), 815–8.

Drost, H.G., Gabel, A., Grosse, I. and Quint, M. (2015). Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis. *Molecular biology and evolution 32*(5), 1221–31.

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), 135–142.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America 95*(25), 14863–8.

Endler, J.A. (1986). *Natural Selection in the Wild.* Monographs in population biology. Princeton University Press.

Ernst, J. and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC bioinformatics 7*, 191.

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

Eyre-Walker, A. and Keightley, P.D. (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), 2097–108.

Eyre-Walker, A., Woolfit, M. and Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics 173*(2), 891–900.

Ferson, S., Rohlf, F.J. and Koehn, R.K. (1985). Measuring Shape Variation of Two-Dimensional Outlines. *Systematic Zoology 34*(1), 59.

Filipowicz, W., Jaskiewicz, L., Kolb, F.A. and Pillai, R.S. (2005). Post-transcriptional gene silencing by siRNAs and miRNAs. *Current Opinion in Structural Biology 15*(3), 331–341.

Forgacs, G. and Newman, S.A. (2005). *Biological Physics of the Developing Embryo.* Cambridge University Press.

Frise, E., Hammonds, A.S. and Celniker, S.E. (2010). Systematic image-driven analysis of the spatial Drosophila embryonic expression landscape. *Molecular systems biology 6*, 345.

Gall, J.G. and Pardue, M.L. (1969). Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America 63*(2), 378–383.

Garcia-Bellido, A. and Merriam, J.R. (1969). Cell lineage of the imaginal discs in Drosophila gynandromorphs. *Journal of Experimental Zoology 170*(1), 61–75.

García-Bellido, A., Ripoll, P. and Morata, G. (1973). Developmental Compartmentalisation of the Wing Disk of Drosophila. *Nature 245*(147), 251–253.

Gelbart, Emmert, Gelbart, W.M. and Emmert, D.B. (2013). FlyBase High Throughput Expression Pattern Data.

Ghiselin, M.T. (2005). Homology as a relation of correspondence between parts of individuals. *Theory in biosciences = Theorie in den Biowissenschaften 124*(2), 91–103.

Gilbert, S.F. (1998). Conceptual breakthroughs in developmental biology. *Journal of Bio-*

sciences 23(3), 169–176.

Gilbert, S.F. (2007). Fate maps, gene expression maps, and the evidentiary structure of evolutionary developmental biology. In J. Maienschein and M. D. Laubichler (Eds.), *From Embryology to Evo-Devo : A History of Developmental Evolution.*, Dibner Institute Studies in the History of Science and Technology, pp. 357–374. The MIT Press.

Gilbert, S.F. (2011). Expanding the Temporal Dimensions of Developmental Biology: The Role of Environmental Agents in Establishing Adult-Onset Phenotypes. *Biological Theory 6*(1), 65–72.

Gilbert, S.F. (2014). *Developmental Biology* (10th ed.). Sinuaer Associates.

Gillespie, J.H. (1994). *The Causes of Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press.

Gloor, G.B., Martin, L.C., Wahl, L.M. and Dunn, S.D. (2005). Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions. *Biochemistry 44*(19), 7156–7165.

Gohlmann, H. and Talloen, W. (2009). *Gene Expression Studies Using Affymetrix Microarrays*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press.

Graveley, B.R. et al (2011). The developmental transcriptome of Drosophila melanogaster. *Nature 471*(7339), 473–9.

Griesemer, J. (2014). Reproduction and scaffolded developmental processes: an integrated evolutionary perspective. In *Towards a Theory of Development*, pp. 183–202. Oxford University Press.

Grzybowska, E.A., Wilczynska, A. and Siedlecki, J.A. (2001). Regulatory Functions of 3 UTRs. *Biochemical and Biophysical Research Communications 288*(2), 291–295.

Gunz, P., Mitteroecker, P. and Bookstein, F.L. (2005). Semilandmarks in Three Dimensions. In *Modern Morphometrics in Physical Anthropology*, pp. 73–98. New York: Kluwer Academic Publishers-Plenum Publishers.

Gurunathan, R., Van Emden, B., Panchanathan, S. and Kumar, S. (2004). Identifying spatially similar gene expression patterns in early stage fruit fly embryo images: binary feature versus invariant moment digital representations. *BMC bioinformatics 5*, 202.

Haeckel, E. (1874). *Anthropogenie oder Entwickelungsgeschichte des Menschen*. Engelmann, Leipzig.

Haeckel, E. (1880). *The history of creation*, Volume 1. New York: Appleton and Company.

Haeckel, E. (1903). *Anthropogenie: oder, Entwickelungsgeschichte des menschen* (5th ed.). Leipzig: W. Engelmann.

Haerty, W. et al (2007). Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. *Genetics 177*(3), 1321–35.

Hahn, M.W. and Wray, G.A. (2002). The g-value paradox. *Evolution and Development 4*(2), 73–75.

Hammonds, A.S. et al (2013). Spatial expression of transcription factors in Drosophila embryonic organ development. *Genome biology 14*(12), R140.

Hannenhalli, S.S. and Russell, R.B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *Journal of Molecular Biology 303*(1), 61–76.

Hartenstein, V. (1993). *Atlas of Drosophila Development*. Cold Spring Harbor Laboratory Press.

Hashimshony, T., Feder, M., Levin, M., Hall, B.K. and Yanai, I. (2014). Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature advance on.*

Henry, J.J. (2002). Conserved Mechanism of Dorsoventral Axis Determination in Equal-Cleaving Spiralians. *Developmental Biology 248*(2), 343–355.

Hereford, J., Hansen, T.F. and Houle, D. (2004). Comparing strengths of directional selection: how strong is strong? *Evolution; international journal of organic evolution 58*(10), 2133–43.

Hershberg, R. and Petrov, D.A. (2008). Selection on codon bias. *Annual review of genetics 42*, 287–99.

Hoekstra, H.E. et al (2001). Strength and tempo of directional selection in the wild. *Proceedings of the National Academy of Sciences of the United States of America 98*(16), 9157–60.

Hooper, S.D. et al (2007). Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. *Molecular systems biology 3*, 72.

Horder, T. (2010). History of Developmental Biology. In *Encyclopedia of Life Sciences*. Chichester, UK: John Wiley & Sons, Ltd.

Hotta, K. et al (2007). A web-based interactive developmental table for the ascidian Ciona intestinalis, including 3D real-image embryo reconstructions: I. From fertilized egg to hatching larva. *Developmental dynamics : an official publication of the American Association of Anatomists 236*(7), 1790–805.

Huang, W. et al (2014). Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome research 24*(7), 1193–208.

Hudson, C., Lotito, S. and Yasuo, H. (2007). Sequential and combinatorial inputs from Nodal, Delta2/Notch and FGF/MEK/ERK signalling pathways establish a grid-like organisation of distinct cell identities in the ascidian neural plate. *Development 134*(19), 3527–3537.

Huxley, J. and De Beer, G. (1963). *The elements of experimental embryology*. Cambridge com-

parative physiology. Hafner Pub. Co.

I. L. Dryden, K.V.M. (1998). *Statistical Shape Analysis* (1 ed.). Wiley.

Imai, K.S. (2003). A Twist-like bHLH gene is a downstream factor of an endogenous FGF and determines mesenchymal fate in the ascidian embryos. *Development 130*(18), 4461–4472.

Imai, K.S., Hino, K., Yagi, K., Satoh, N. and Satou, Y. (2004). Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. *Development (Cambridge, England) 131*(16), 4047–58.

Imai, K.S., Levine, M., Satoh, N. and Satou, Y. (2006). Regulatory blueprint for a chordate embryo. *Science (New York, N.Y.) 312*(5777), 1183–7.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology 3*, 318–56.

Jaeger, J. (2011). The gap gene network. *Cellular and molecular life sciences : CMLS 68*(2), 243–74.

Jaeger, J. and Sharpe, J. (2014). On the concept of mechanism in development. In A. Minelli and T. Pradeu (Eds.), *Towards a Theory of Development*, pp. 56–78. OUP Oxford.

James Rohlf, F. and Marcus, L.F. (1993). A revolution morphometrics. *Trends in Ecology & Evolution 8*(4), 129–132.

Janning, W. (1978). Gynandromorph Fate Maps in Drosophila. In W. J. Gehring (Ed.), *Results and Problems in Cell Differentiation*, Chapter 1, pp. 1–28. Springer-Verlag Berlin Heidelberg.

Jernvall, J. and Selänne, L. (1999). Laser confocal microscopy and geographic information systems in the study of dental morphology. *Palaeontologia electronica 2*(1), 18.

Jolicoeur, P. and Mosimann, J.E. (1960). Size and shape variation in the painted turtle. A principal component analysis. *Growth 24*, 339–54.

Kalinka, A.T. and Tomancak, P. (2012). The evolution of early animal embryos: conservation or divergence? *Trends in ecology & evolution 27*(7), 385–93.

Kalinka, A.T. et al (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature 468*(7325), 811–4.

Kang, T.S. and Kini, R.M. (2009). Structural determinants of protein folding. *Cellular and Molecular Life Sciences 66*(14), 2341–2361.

Kapp, L.D. and Lorsch, J.R. (2004). The Molecular Mechanics of Eukaryotic Translation. *Annual Review of Biochemistry 73*(1), 657–704.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature 423*(6937), 241–254.

Kimchi-Sarfaty, C. et al (2007). A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science (New York, N.Y.) 315*(5811), 525–8.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature 217*(5129), 624–6.

Kimura, M. (1969). The rate of molecular evolution considered from the standpoint of population genetics. *Proceedings of the National Academy of Sciences 63*(4), 1181–1188.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kimura, M. and Ohta, T. (1974). On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America 71*(7), 2848–52.

King, J.L. and Jukes, T.H. (1969). Non-Darwinian evolution. *Science (New York, N.Y.) 164*(3881), 788–98.

Klingenberg, C.P. (2010). Evolution and development of shape: integrating quantitative approaches. *Nature Reviews Genetics 11*(9), 623–35.

Konikoff, C.E., Karr, T.L., McCutchan, M., Newfeld, S.J. and Kumar, S. (2012). Comparison of embryonic expression within multigene families using the FlyExpress discovery platform reveals more spatial than temporal divergence. *Developmental dynamics : an official publication of the American Association of Anatomists 241*(1), 150–60.

Kozak, M. (1992). Regulation of Translation in Eukaryotic Systems. *Annual Review of Cell Biology 8*(1), 197–225.

Kuhl, F.P. and Giardina, C.R. (1982). Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing 18*(3), 236–258.

Kumar, S. et al (2011). FlyExpress: visual mining of spatiotemporal patterns for genes and publications in Drosophila embryogenesis. *Bioinformatics (Oxford, England) 27*(23), 3319–20.

Lamarck, J. (1809). *Zoological philosophy*. Univ. of Chicago Press, Chicago.

Lécuyer, E. et al (2007). Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. *Cell 131*(1), 174–187.

Lemaire, P. (2009). Unfolding a chordate developmental program, one cell at a time: invariant cell lineages, short-range inductions and evolutionary plasticity in ascidians. *Developmental biology 332*(1), 48–60.

Lemaire, P., Smith, W.C. and Nishida, H. (2008). Ascidians and the plasticity of the chordate developmental program. *Current biology : CB 18*(14), R620–31.

Levin, M., Hashimshony, T., Wagner, F. and Yanai, I. (2012). Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo. *Developmental Cell 22*(5), 1101–1108.

Lichtenstein, F., Antoneli, F. and Briones, M.R.S. (2015). MIA: Mutual Information Analyzer, a graphic user interface program that calculates entropy, vertical and horizontal mutual information of molecular sequence sets. *BMC Bioinformatics 16*(1), 409.

Lohmann, G.P. (1983). Eigenshape analysis of microfossils: A general morphometric procedure for describing changes in shape. *Journal of the International Association for Mathematical Geology 15*(6), 659–672.

Longo, G., Miquel, P.A., Sonnenschein, C. and Soto, A. (2012). Is information a proper observable for biological organization? *Progress in Biophysics and Molecular Biology 109*(3), 108–114.

Mackay, T.F.C. et al (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature 482*(7384), 173–8.

Mann, M. and Jensen, O.N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology 21*(3), 255–261.

Marianes, A. and Spradling, A.C. (2013). Physiological and stem cell compartmentalization within the Drosophila midgut. *eLife 2*, e00886.

Martínez-Abadías, N., Mateu, R., Niksic, M., Russo, L. and Sharpe, J. (2016). Geometric Morphometrics on Gene Expression Patterns Within Phenotypes: A Case Example on Limb Development. *Systematic biology 65*(2), 194–211.

Matsuoka, T., Ikeda, T., Fujimaki, K. and Satou, Y. (2013). Transcriptome dynamics in early embryos of the ascidian, Ciona intestinalis. *Developmental biology 384*(2), 375–85.

Mayr, E. (1997). *Evolution and the Diversity of Life: Selected Essays*. Selected Essays. Belknap Press of Harvard University Press.

McDonald, J.H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature 351*(6328), 652–4.

McLellan, T. and Endler, J.A. (1998). The Relative Success of Some Methods for Measuring and Describing the Shape of Complex Objects. *Systematic Biology 47*(2), 264–281.

McShea, D.W. (1996). Perspective: Metazoan Complexity and Evolution: Is There a Trend? *Evolution 50*(2), 477.

McShea, D.W. (2015). Three Trends in the History of Life: An Evolutionary Syndrome. *Evolutionary Biology*.

Medawar, P. (1954). The significance of inductive relationships in the development of vertebrates. *Journal of Embryology and Experimental ... 2*(June), 172–174.

Mensch, J., Serra, F., Lavagnino, N.J., Dopazo, H. and Hasson, E. (2013). Positive selection in nucleoporins challenges constraints on early expressed genes in Drosophila development. *Genome biology and evolution 5*(11), 2231–41.

Messer, P.W. and Petrov, D.A. (2013). Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America 110*(21), 8615–20.

Minelli, A. (2011). Animal Development, an Open-Ended Segment of Life. *Biological Theory 6*(1), 4–15.

Minelli, A. (2014). Developmental disparity. In *Towards a Theory of Development*, pp. 227–245. Oxford University Press.

Mitani, Y., Takahashi, H. and Satoh, N. (1999). An ascidian T-box gene As-T2 is related to the Tbx6 subfamily and is associated with embryonic muscle cell differentiation. *Developmental Dynamics 215*(1), 62–68.

Mitteroecker, P. and Gunz, P. (2009). Advances in Geometric Morphometrics. *Evolutionary Biology 36*(2), 235–247.

Moczek, A.P. (2014). Towards a theory of development through a theory of developmental evolution. In *Towards a Theory of Development*, pp. 218–226. Oxford University Press.

Monod, J. (1963). Genetic Repression, Allosteric Inhibition, and Cellular Differentiation. In M. Locke (Ed.), *Cytodifferentiation and Macromolecular Synthesis*, pp. 30–64. New York: Academic Press.

Morgan, T.H. (1919). The physical basis of heredity.

Mukai, T. (1964). THE GENETIC STRUCTURE OF NATURAL POPULATIONS OF DROSOPHILA MELANOGASTER. I. SPONTANEOUS MUTATION RATE OF POLYGENES CONTROLLING VIABILITY. *Genetics 50*, 1–19.

Nakamura, M.J., Terai, J., Okubo, R., Hotta, K. and Oka, K. (2012). Three-dimensional anatomy of the Ciona intestinalis tailbud embryo at single-cell resolution. *Developmental biology 372*(2), 274–84.

Nicol, D. and Meinertzhagen, I.A. (1988). Development of the central nervous system of the larva of the ascidian, Ciona intestinalis L. II. Neural plate morphogenesis and cell lineages during neurulation. *Developmental biology 130*(2), 737–66.

Nishida, H. (1987). Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme. III. Up to the tissue restricted stage. *Developmental biology 121*(2), 526–41.

Nishida, H. (2005). Specification of embryonic axis and mosaic development in ascidians. *Developmental dynamics : an official publication of the American Association of Anatomists 233*(4), 1177–93.

Nuzhdin, S.V., Wayne, M.L., Harmon, K.L. and McIntyre, L.M. (2004). Common pattern of evolution of gene expression level and protein

sequence in Drosophila. *Molecular biology and evolution 21*(7), 1308–17.

OHTA, T. (1973). Slightly Deleterious Mutant Substitutions in Evolution. *Nature 246*(5428), 96–98.

Ohta, T. and Gillespie, J.H. (1996). Development of Neutral and Nearly Neutral Theories. *Theoretical Population Biology 49*(2), 128–142.

Ostrowski, S., Dierick, H.A. and Bejsovec, A. (2002). Genetic Control of Cuticle Formation During Embryonic Development of Drosophila melanogaster. *Genetics 161*(1), 171–182.

Oyama, S. (2000). *The Ontogeny of Information: Developmental Systems and Evolution*. Science and Cultural Theory. Duke University Press.

Papaioannou, V.E. (2014). The T-box gene family: emerging roles in development, stem cells and cancer. *Development (Cambridge, England) 141*(20), 3819–33.

Pavlidis, T. (1980). Algorithms for shape analysis of contours and waveforms. *IEEE transactions on pattern analysis and machine intelligence 2*(4), 301–12.

Peregrín-Alvarez, J.M., Sanford, C. and Parkinson, J. (2009). The conservation and evolutionary modularity of metabolism. *Genome biology 10*(6), R63.

Piasecka, B., Lichocki, P., Moretti, S., Bergmann, S. and Robinson-Rechavi, M. (2013). The hourglass and the early conservation models–coexisting patterns of developmental constraints in vertebrates. *PLoS genetics 9*(4), e1003476.

Plotkin, J.B. and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics 12*(1), 32–42.

Poe, S. and Wake, M.H. (2004). Quantitative tests of general models for the evolution of development. *The American naturalist 164*(3), 415–22.

Pollet, N., Delius, H. and Niehrs, C. (2003). In situ analysis of gene expression in Xenopus embryos. *Comptes rendus biologies 326*(10-11), 1011–1017.

Pool, J.E. et al (2012). Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. *PLoS Genetics 8*(12), e1003080.

Poulin, R. (2011). *Evolutionary Ecology of Parasites: (Second Edition)*. Princeton University Press.

Pradeu, T. (2014). Regenerating theories in developmental biology. In A. Minelli and T. Pradeu (Eds.), *Towards a Theory of Development*, pp. 15–32. Oxford University Press.

Presnyak, V. et al (2015). Codon Optimality Is a Major Determinant of mRNA Stability. *Cell 160*(6), 1111–1124.

Pröschel, M., Zhang, Z. and Parsch, J. (2006). Widespread adaptive evolution of Drosophila genes with sex-biased expression. *Genetics 174*(2), 893–900.

Raff, R.A. (1996). *The Shape of Life: Genes, Development, and the Evolution of Animal Form*. University of Chicago Press, Chicago.

Richardson, M.K. (1995). Heterochrony and the phylotypic period. *Developmental biology 172*(2), 412–21.

Richardson, M.K. et al (1997). There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anatomy and embryology 196*(2), 91–106.

Richardson, M.K. and Keuck, G. (2002). Haeckel's ABC of evolution and development. *Biological Reviews of the Cambridge Philosophical Society 77*(4), 495 – 528.

Roberts, D.B. (1998). *Drosophila: a practical approach*. Practical approach series. IRL Press at Oxford University Press.

Russell, E.S. (1916). *Form and function : a contribution to the history of animal morphology*. London: J. Murray.

Salazar-Ciudad, I. (2010). Morphological evolution and embryonic developmental diversity in metazoa. *Development (Cambridge, England) 137*(4), 531–9.

Salazar-Ciudad, I., Newman, S.A. and Solé, R.V. (2001). Phenotypic and dynamical transitions in model genetic networks. I. Emergence of patterns and genotype-phenotype relationships. *Evolution & development 3*(2), 84–94.

Salvador-Martínez, I. and Salazar-Ciudad, I. (2015). How complexity increases in development: An analysis of the spatial-temporal dynamics of 1218 genes in Drosophila melanogaster. *Developmental Biology 405*(2), 328–339.

Sander, K. (1983). The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In B. C. Goodwin, N. Holder, and C. C. Wylie (Eds.), *Development and Evolution*. Cambridge, UK: Cambridge University Press.

Sander, K. (1996). Pattern formation in insect embryogenesis: The evolution of concepts and mechanisms. *International Journal of Insect Morphology and Embryology 25*(4), 349–367.

Sanjuán, R., Moya, A. and Elena, S.F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America 101*(22), 8396–401.

Satou, Y., Kawashima, T., Shoguchi, E., Nakayama, A. and Satoh, N. (2005). An integrated database of the ascidian, Ciona intestinalis: towards functional genomics. *Zoological science 22*(8), 837–43.

Satou, Y. et al (2008). Improved genome assembly and evidence-based global gene model set

for the chordate Ciona intestinalis: new insight into intron and operon populations. *Genome biology 9*(10), R152.

Sebé-Pedrós, A. et al (2013). Early evolution of the T-box transcription factor family. *Proceedings of the National Academy of Sciences of the United States of America 110*(40), 16050–5.

Sempere, L.F., Cole, C.N., McPeek, M.A. and Peterson, K.J. (2006). The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *Journal of experimental zoology. Part B, Molecular and developmental evolution 306*(6), 575–88.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 27*(3), 379–423.

Sharp, P.M. (1991). Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution. *Journal of molecular evolution 33*(1), 23–33.

Sharpe, J. (2003). Optical projection tomography as a new tool for studying embryo anatomy. *Journal of Anatomy 202*(2), 175–181.

Showell, C., Binder, O. and Conlon, F.L. (2004). T-box genes in early embryogenesis. *Developmental dynamics : an official publication of the American Association of Anatomists 229*(1), 201–18.

Simpson, G.G. (1953). *The Major Features of Evolution*. New York: Columbia University Press.

Slack, J.M., Holland, P.W. and Graham, C.F. (1993). The zootype and the phylotypic stage. *Nature 361*(6412), 490–2.

Slice, D.E. (Ed.) (2005). *Modern Morphometrics in Physical Anthropology*. Developments in Primatology: Progress and Prospects. New York: Kluwer Academic Publishers-Plenum Publishers.

Smith, J.M. (2000). The Concept of Information in Biology. *Philosophy of Science 67*(2), 177–194.

Stahl, P.L. et al (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science 353*(6294), 78–82.

Stapleton, M. et al (2002). The Drosophila Gene Collection: Identification of Putative Full-Length cDNAs for 70% of D. melanogaster Genes. *Genome Research 12*(8), 1294–1300.

Summerhurst, K., Stark, M., Sharpe, J., Davidson, D. and Murphy, P. (2008). 3D representation of Wnt and Frizzled gene expression patterns in the mouse embryo at embryonic day 11.5 (Ts19). *Gene expression patterns : GEP 8*(5), 331–48.

Tassy, O., Daian, F., Hudson, C., Bertrand, V. and Lemaire, P. (2006). A quantitative approach to the study of cell shapes and interactions during early chordate embryogenesis. *Current biology : CB 16*(4), 345–58.

Tassy, O. et al (2010). The ANISEED database: digital representation, formalization, and elucidation of a chordate developmental program. *Genome research 20*(10), 1459–68.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature 467*(7319), 1061–1073.

Tokuoka, M., Imai, K.S., Satou, Y. and Satoh, N. (2004). Three distinct lineages of mesenchymal cells in Ciona intestinalis embryos demonstrated by specific gene expression. *Developmental biology 274*(1), 211–24.

Tomancak, P. et al (2002). Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biology 3*(12), 1–14.

Tomancak, P. et al (2007). Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome biology 8*(7), R145.

Valentine, J.W., Collins, A.G. and Meyer, C.P. (1994). Morphological Complexity Increase in Metazoans. *Paleobiology 20*(2), 131–142.

von Baer, K.E. (1828). *Über Entwickelungsgeschichte der Thiere. Beobachtung und Reflexion*. Königsberg.

Wallberg, A. et al (2014). A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee Apis mellifera. *Nature Genetics 46*(10), 1081–1088.

Weiszmann, R., Hammonds, A.S. and Celniker, S.E. (2009a). Determination of gene expression patterns using high-throughput RNA in situ hybridization to whole-mount Drosophila embryos. *Nature protocols 4*(5), 605–18.

Weiszmann, R., Hammonds, A.S. and Celniker, S.E. (2009b). Determination of gene expression patterns using high-throughput RNA in situ hybridization to whole-mount Drosophila embryos. *Nature Protocols 4*(5), 605–618.

Wen, X. et al (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences of the United States of America 95*(1), 334–9.

Winchester, J.M. (2016). MorphoTester: An Open Source Application for Morphological Topographic Analysis. *PloS one 11*(2), e0147649.

Winchester, J.M. et al (2014). Dental topography of platyrrhines and prosimians: Convergence and contrasts. *American Journal of Physical Anthropology 153*(1), 29–44.

Wolfram, S. (2002). *A new kind of science*. Wolfram Media.

Wray, G.A. (2000). The evolution of embryonic patterning mechanisms in animals. *Seminars in cell & developmental biology 11*(6), 385–93.

Xia, X. (1996). Maximizing transcription efficiency causes codon usage bias. *Genetics 144*(3), 1309–20.

Yasuo, H. and Satoh, N. (1998). Conservation of the developmental role of Brachyury in noto-

chord formation in a urochordate, the ascidian Balocynthia roretzi. *Developmental biology 200*(2), 158–70.

Yockey, H.P. (2005). *Information Theory, Evolution, and the Origin of Life*. Cambridge University Press.

Zelditch, M.L., Swiderski, D.L. and Sheets, H.D. (2012). *Geometric Morphometrics for Biologists: A Primer*. Academic Press. Elsevier Academic Press.

Zelditch, M.L., Wood, A.R., Bonett, R.M. and Swiderski, D.L. (2008). Modularity of the rodent mandible: Integrating bones, muscles, and teeth. *Evolution & Development 10*(6), 756–768.