

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

Copyright © 2016 Irepan Salvador-Martínez

ISSN 1238-8645

ISBN 000-00-0000-0 (paperback)

ISBN 000-00-0000-0 (PDF)

Helsinki 2016

Unigrafia

--

<h1>Acknowledgements</h1>	
---------------------------	--

Contents

List of publications

Abbreviations

Abstract

1	Review of the literature	1
1.1	What is development?	2
1.2	Complexity	2
1.2.1	Different definitions of complexity	3
1.2.2	Complexity Increase in Development	6
1.2.3	Spatial complexity	9
1.3	Adaptation	12
1.3.1	Molecular evolution	13
1.3.2	Neutral theory of evolution	14
1.3.3	From neutral to nearly neutral theories	14
1.3.4	McDonald-Kreitman test	15
1.3.5	Distribution of Fitness Effects	16
1.3.6	DFE-alpha method	16
1.3.7	The <i>Drosophila melanogaster</i> Genetic Reference Panel	18
1.4	<i>Drosophila</i>	18
1.4.1	<i>D. melanogaster</i> life cycle	20
1.4.2	Fate map	20
1.4.3	Gene expression databases of <i>D. melanogaster</i>	22
1.5	The Hourglass model in <i>Drosophila</i>	23
1.6	<i>Ciona</i>	25
1.6.1	<i>Ciona</i> as a model	25
1.6.2	Current knowledge about <i>Ciona</i> development	25
1.6.3	<i>Ciona intestinalis</i> life cycle	25
2	Aims of the study	27
3	Material and Methods	28
4	Results and Discussion	29
4.1	Comparative study between <i>Drosophila</i> and <i>Ciona</i> (I and II)	29
4.1.1	Compartmentalization	29
4.1.2	Disparity	30
4.1.3	The leading role of TFs and GFs (and other signalling molecules)	31
4.1.4	2D and 3D roughness analyses	32
4.1.5	Synexpression territories	33
4.2	Main spatio-temporal profiles of gene expression in <i>Drosophila</i> (I)	36
4.3	Discrepancies between fate map and STs (II)	36
4.4	Adaptation in <i>Drosophila</i> embryogenesis (III and IV)	37
4.4.1	STs or anatomical terms with high ω_α	37
4.4.2	STs or anatomical terms with low ω_α	38
4.4.3	Transcriptome age index and other genomic determinants	39
4.4.4	Selective constraint in late embryogenesis	40
4.5	Adaptation trough <i>Drosophila</i> life cycle (IV)	40
5	Concluding Remarks	43
	References	44

<h2>List of publications</h2>	

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 <i>Drosophila melanogaster</i> 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

Contents

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 is 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 situ 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 developmental biology. Since the 1980's crucial discoveries (Gilbert, 1998) have not only improved our understanding of the developmental process, but also changed the perspective of the explanatory role of development in biology.

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, although having increased substantially the knowledge in the developmental biology field, is often not suited to gain a global view that could serve to answer interesting general questions in Biology. In order to attain this global view a systemic statistical approach would be fruitful.

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 systemic 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 1.5) supported by population genomics data?

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 genetics. Nowadays, the combination of these scientific fields form multiple research programmes. Indeed, modern evolutionary developmental biology (evo-devo) is the explicit combination of the first two fields.

However, these fields have not always gone hand in hand. Some decades ago, there was a clear conceptual and epistemological separation between evolutionary biology (mostly practised by geneticists) and developmental biology, even though embryology (which slowly transformed into developmental biology in the middle of the 20th century, see Horder, 2010) was considered crucial for the study of evolution in the 19th century.

In the following section, I will give a brief introduction of the scientific and philosophical origins of developmental biology, with special attention to its relations with evolutionary biology and genetics (for a comprehensive review on this issue, see: Amundson, 2005; Gilbert, 1991), and to some of the concepts I will use in this dissertation.

Before that, it might be useful to define what development is, so first, I will address this apparently simple question.

1.1 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.

1.2 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. I will first try to explain different existing definitions of organismic complexity.

Then, I will explore the relationship between complexity in evolution and development, and discuss the possibility of a trend in terms of complexity increase.

1.2.1 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, \dots, 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:

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

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 $j = 1, \dots, L$ 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 thought 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 has restricted 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). It has also been noted (Jaeger and Sharpe, 2014) that the use of the informational approach is not appropriate to the study of development, simply because Shannon's information can not increase while it is transmitted. If applied to development, it would imply that the amount of information (or complexity) of the egg and the adult remain constant (Jaeger and Sharpe, 2014), which contradicts the generalized notion that the complexity increases during embryogenesis.

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).

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). Even when in his example it would seem easy to discern a simple GRN from a complex one just from its topology, the high intricacy of real biological GRN would make their distinction (at least by visual inspection) a very difficult task.

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). Importantly, with this definition, complexity at different levels are not necessarily correlated.

This lack of correspondence at different levels using "the number of part types" is 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.

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 if it has more gene-gene interactions than the other.

1.2.2 Complexity Increase in Development

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).

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 when a new cell type has formed during development. In addition, this definition does not take into account that embryos do not only get more cell types, but they also get organized in specific patterns in space and time in the embryo, which also could be considered as an increase in complexity over developmental time.

The notion of an increase in complexity during embryogenesis is tightly related to the concept of embryo compartmentalization.

Box1. On the relationship between the increase of complexity in Evolution and Development

The connection between the increase in complexity during development and evolutionary time it 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 section X).

This implies 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). 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 1.3).

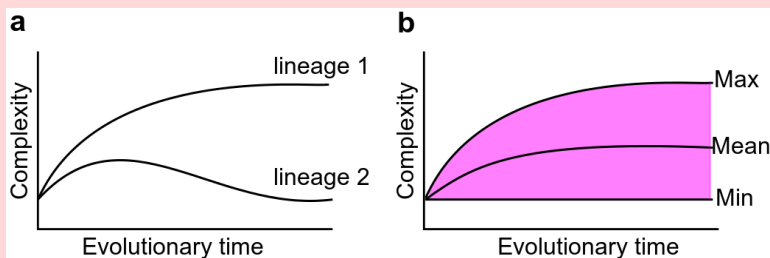


Figure 1.3: Two lineages with different complexity change through their evolutionary trajectories. b) Representation of the minimum, mean and maximum complexity of many lineages over evolutionary time in which the minimum stay constant while the mean and maximum increase. Redrawn from (Arthur, 2010).

If we consider unicellularity 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.

If we look at maximum complexity a trend for increasing complexity would be apparent, as the initial 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.

It is important to notice that this apparent trend towards increasing complexity does not necessarily imply that it has been selected for (McShea, 2015).

If complexity would follow a random walk scenario starting from unicellular organisms (in a random walk the mean distance of a point after n time steps is $\sqrt[2]{n}$), the mean and maximum would show a trend towards increasing complexity 1.3 given that there is a minimum complexity requirement (it is not possible to have less than one cell) (Gould, 1996).

Compartmentalization in development

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 García Bellido 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

1.2 Complexity

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 (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.

1.2.3 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). In order to analyse the shape of the internal parts of an embryo, staining and imaging methods are usually required. One useful approach is the use of techniques that highlight the distribution of a tissue-specific gene products (e.g., whole-mount in situ mRNA hybridization or immuno-histochemistry techniques). Therefore, the shape of gene expression patterns (represented in 2D or 3D) can be used to describe the complexity of the embryo. When analyzing the shape of a gene expression through development, it could be that the expression pattern: 1) does not change its complexity, 2) becomes more complex, or 3) becomes less complex. However, it would be expected that when analysing a large number of genes, the shape complexity would increase at a statistical level.

In order to analyse the complexity of the shape of an object is necessary to quantitatively measure and analyse it. 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 (Jolicœur 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 analytical 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 used to quantify teeth surfaces as if they were landscapes (Jernvall and Sel  nne, 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.

Gene expression shape analyses

Gene expression patterns can be represented in 2D as a closed outline or as a 3D surface. It is important to mention that a gene expression pattern reflects necessarily the spatial distribution of the cells expressing such gene. When analysing and comparing the shape of many different 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. Therefore, a landmark-free method (like outline methods for 2D or surface methods for 3D data) would be ideal to deal with this type of data.

There are practically no studies in the literature that have objectively and accurately quantified and compared gene expression patterns in a systematic manner (one exception is the recent study of Mart  nez-Abad  as 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 (see methods and study I). In order to use a similar measure of curvature in 3D, I used the Dirichlet normal energy (DNE) described above which quantifies the

deviation of a surface from being planar (see methods and study II). Importantly, both measures are normalized to remove size and orientation effects.

1.3 Adaptation

Measuring adaptation is a central theme in evolutionary biology. In many texts however, adaptation is not clearly defined (leading to ambiguity) or is often used in biological irrelevant contexts (Dobzhansky, 1968). Even in Darwin's *Origin*, where it is a central concept, adaptation is not explicitly defined throughout the text (Darwin, 1859).

Adaptation

Usually adaptation can refer to two different things, to a trait that enables or enhances the probability of its bearer surviving or reproducing in a given environment (Dobzhansky, 1956) and to the process to become adapted. Simpson (1953) said that:

"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 Simpson)

We say an organism is adapted to an environment when it is able to live and reproduce in it. A related concept is adaptedness, or "the degree to which an organism is adapted to an environment" (Dobzhansky, 1968).

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 Darwinian 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.

1.3 Adaptation

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 (REF Darwin), 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.

Among the latter 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.

1.3.1 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 (REF)). 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 (nonsense 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).

1.3.2 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 (REF).

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.

1.3.3 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. One of the most popular tests 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.

1.3.4 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). Mutations under positive selection therefore would be expected to spread through a population rapidly so they would not contribute to polymorphism but only to divergence substitutions.

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. 1.4). 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.

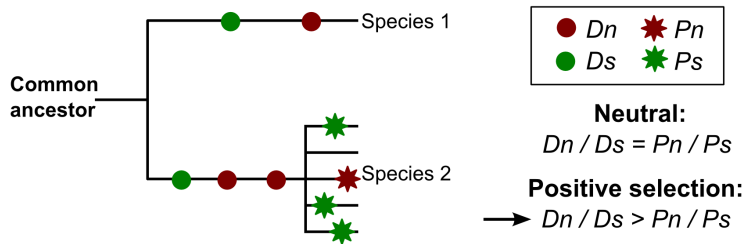


Figure 1.4: McDonald-Kreitman Test (MKT). MKT compares the ratio of non-synonymous (Dn ; red circle) to synonymous (Ds ; green circle) divergence (Dn/Ds) to the ratio of non-synonymous (Pn ; red star) to synonymous (Ps ; green star) polymorphic changes (Pn/Ps). Positive selection is detected when $Dn/Ds > Pn/Ps$, 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.

1.3.5 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 (Eyre-Walker and Keightley, 2007), with weakly deleterious, neutral and slightly adaptive mutations in between.

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).

1.3.6 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 1.5).

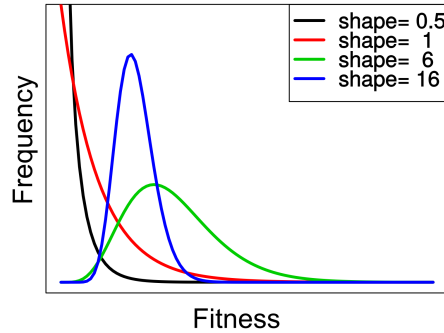


Figure 1.5: Example of different Distribution of Fitness Effects (DFE) represented by a gamma distribution. Many distributions can be represented by modifying the shape parameter of a gamma distribution, from a leptokurtic (shape parameter less than 1) to an exponential (shape parameter equal to 1) or a skewed normal distribution (shape greater than 1).

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 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 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 in the case of genetic draft, background selection or demography changes (Messer and Petrov, 2013).

Inferring natural selection using a population genomics approach

For taking advantage of these new methods that infer adaptation combining polymorphism and divergence data, a populations genomics approach is necessary. In the last years, different 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.

In this study, I used data from the *Drosophila melanogaster* Genetic Reference Panel (DGRP) (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.

1.3.7 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).

1.4 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.

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 underlying 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. Cephalic 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	Stomodaeal plate formation. Cephalic furrow no longer visible.
10	5h 10min to 6h 50min	Stomodaeal 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 yolk 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.

1.4.1 *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.

1.4.2 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., constructing 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

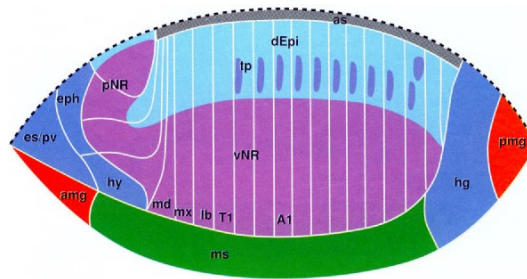


Figure 1.6: Fate map of the *Drosophila melanogaster* blastoderm The fate map is projected onto a planimetric reconstruction of the blastoderm. The upper dashed line represents the dorsal midline and the lower margin represents the ventral midline. A1 Abdominal segment 1; amg anterior midgut rudiment (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)

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 Hörstadius provided the fate map of the sea urchin embryo, also by cell lineage tracking. These two species

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 1.6).

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; Weiszmamm 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 1.6.

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.

1.4.3 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 (Weismann 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 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.

1.5 The Hourglass model in *Drosophila*

As I described briefly in section ??, 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 (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). Salazar-Ciudad (2010)

has proposed that different patterns of variation throughout development in metazoan groups would correlate with different developmental types (a classification based on the relative use of signalling and morphogenetic events).

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 showed that most genes fit best to models incorporating stabilizing selection and proposed that natural selection acts to conserve patterns of gene expression during mid-embryogenesis (Kalinka et al., 2010).

The HG seems also to be reflected in the age of the transcriptome (mid-embryonic stage shows the older transcriptome; Domazet-Lošo and Tautz, 2010) and in the conservation of the regulatory regions (most conserved for genes expressed in mid-development; Piasecka et al., 2013).

Studies measuring the conservation of genes at the DNA sequence level also seem to support the HG. 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 than 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 1.3.6). 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)

1.6 Ciona

1.6.1 Ciona as a model

The ascidian *Ciona intestinalis*, a marine invertebrate animal, has a long history in developmental biology and evolutionary 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).

1.6.2 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 tadpole 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.

1.6.3 *Ciona intestinalis* life cycle

Developmental stages

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

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 2-5	Zygote Early cleavage	0min to 55min 55min to 3h	From the fertilisation event up to the end of the first mitotic cycle 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 tailbud	8.5h to 10h	Clear separation between trunk and tail. Neuropore closure. Tail starts bending
21-22	Mild 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

--

<p>Aims of the study</p>	
--------------------------	--

Material and Methods

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,

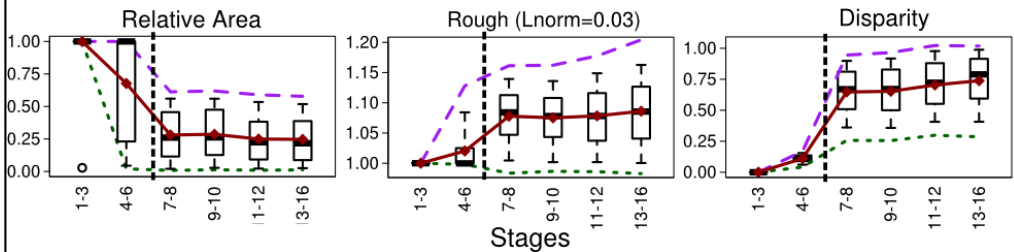


Figure 4.1: Measures in *Drosophila*. Distribution plot of the relative area of expression (left), roughness (center) and disparity (right) for all genes in each stage. Diamonds represent the mean, boxes the Inter Quartile Range (IQR). Whiskers 10 and 90 percentiles. Dashed line represents the max values 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

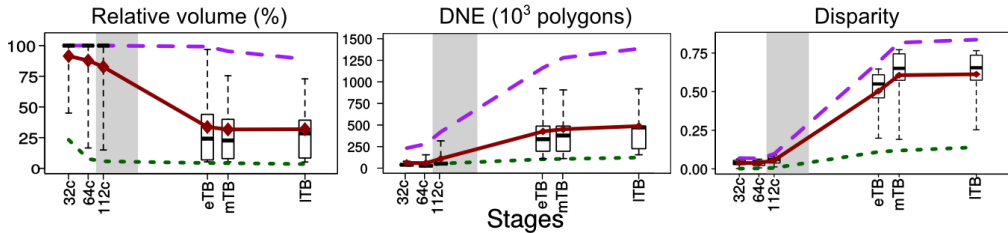


Figure 4.2: Measures in *Ciona*. Distribution plot of the relative volume of expression (left), DNE (center) and disparity (right) for all genes in each stage. Diamonds represent the mean, boxes the IQR. Whiskers 10 and 90 percentiles. Dashed line represents the max values and dotted line the min values (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.

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

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 to 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 GTerm analysis in *Drosophila*, I found that TFs (GO:0003700) and GFs (GO:0008083) showed smaller relative area of expression than 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, TGF β , 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 TGF β 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 TGF β 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

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

the roughness values are significantly higher at smaller spatial scales is significantly higher than 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

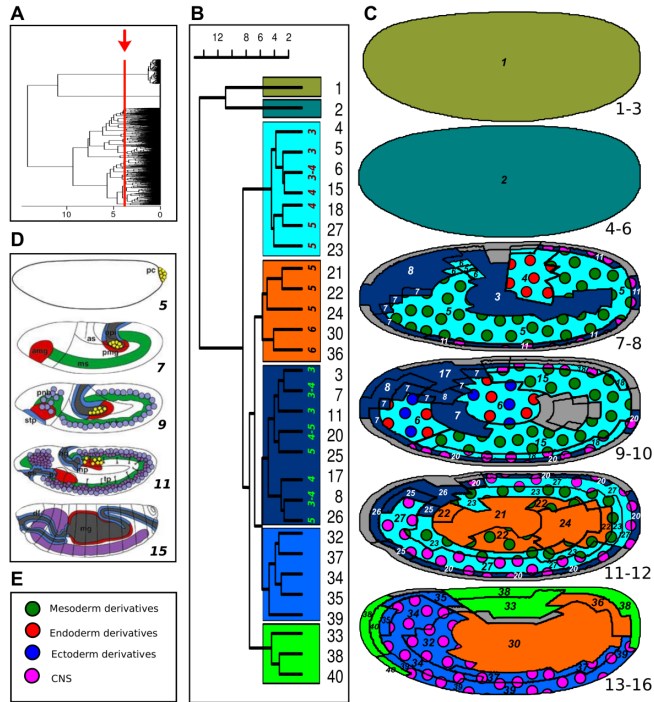


Figure 4.3: Synexpression territories (ST). (A) Dendrogram produced by hierarchical clustering on a similarity matrix (pearson's correlation) of all the embryo regions of the six stages. Red line shows the cut-off to produce 40 STs. (B) Dendrogram reconstructed using only territories with at least 50 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 GO term 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 GO terms 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 GO terms (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 GO terms like 'dendrite morphogenesis' (GO:0048813) and 'axon guidance' (GO:0007411).

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

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

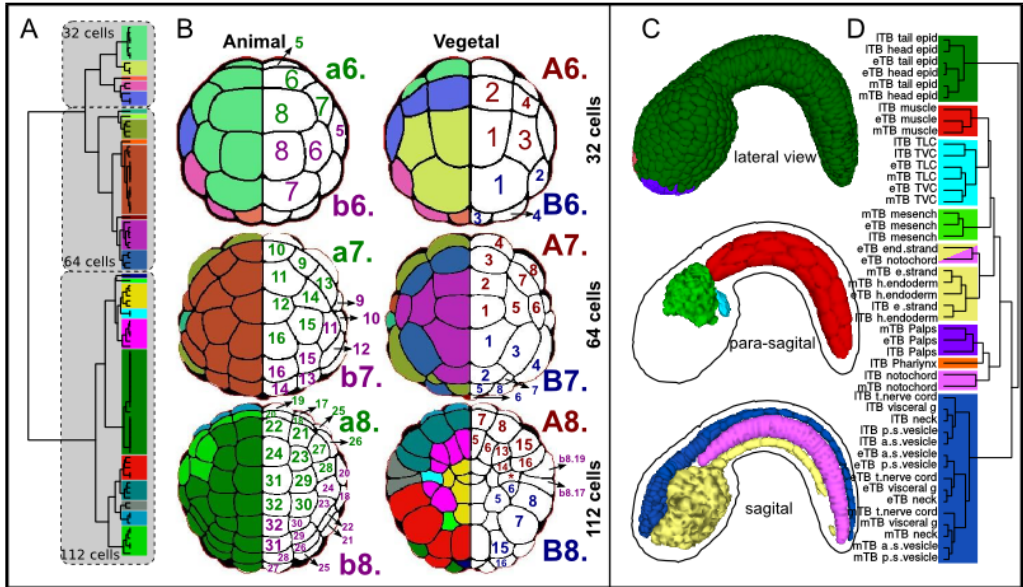


Figure 4.4: *Ciona* synexpression territories. (A) Dendrogram produced by hierarchical clustering of cells in 32-cell, 64-cell and 112-cell stages. Dashed boxes show that STs cluster by stage. Coloured boxes show the cut-off to produce 24 STs. (B) Names of cells (Conklin nomenclature; Conklin, 1905) 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-sagittal and sagittal 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:0016558) GO term 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 were 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 reflects 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).

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

This analysis could not be made in *Drosophila* as the gene expression data is not at 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

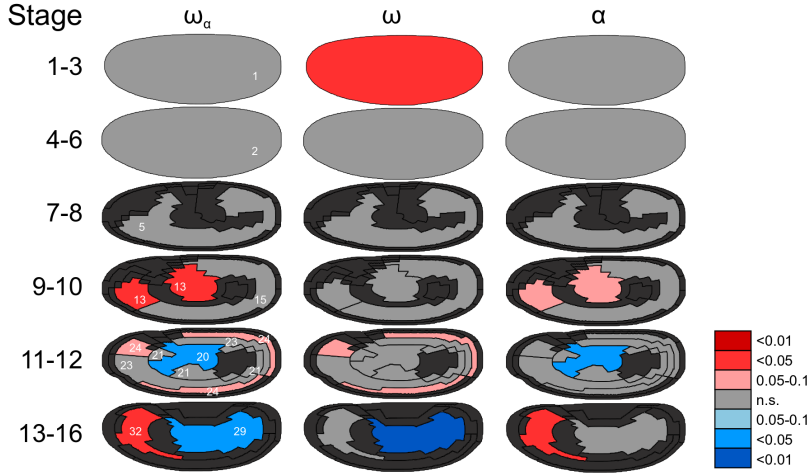


Figure 4.5: ω_α on embryonic territories over space and time. Territories drawn in red in the central column mark significantly high ω_α while those in blue mark significantly low ω_α in space in each of the 6 developmental stages (rows). Other columns depict α , the proportion of base substitutions 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-

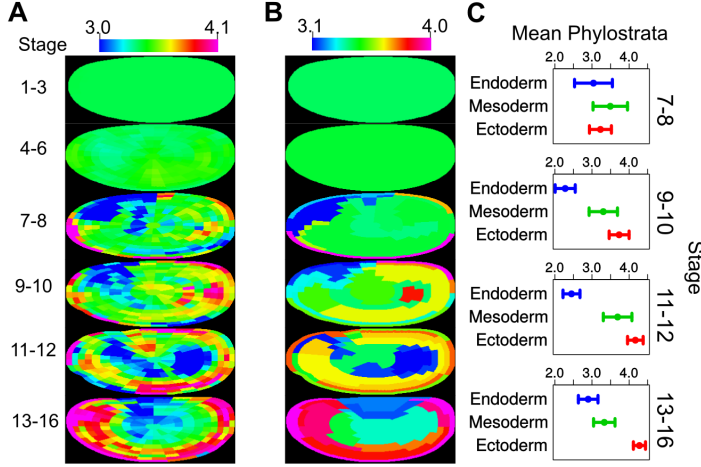


Figure 4.6: The center of the embryo expresses older genes. (A) Heatmaps showing the transcriptome age index (TAI) in polar regions (B) Heatmaps showing the TAI for STs. (C) Mean phylostrata of genes assigned to each germ layer. Circles represent the mean and whiskers the SEM.

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 .

4.5 Adaptation trough *Drosophila* life cycle (IV)

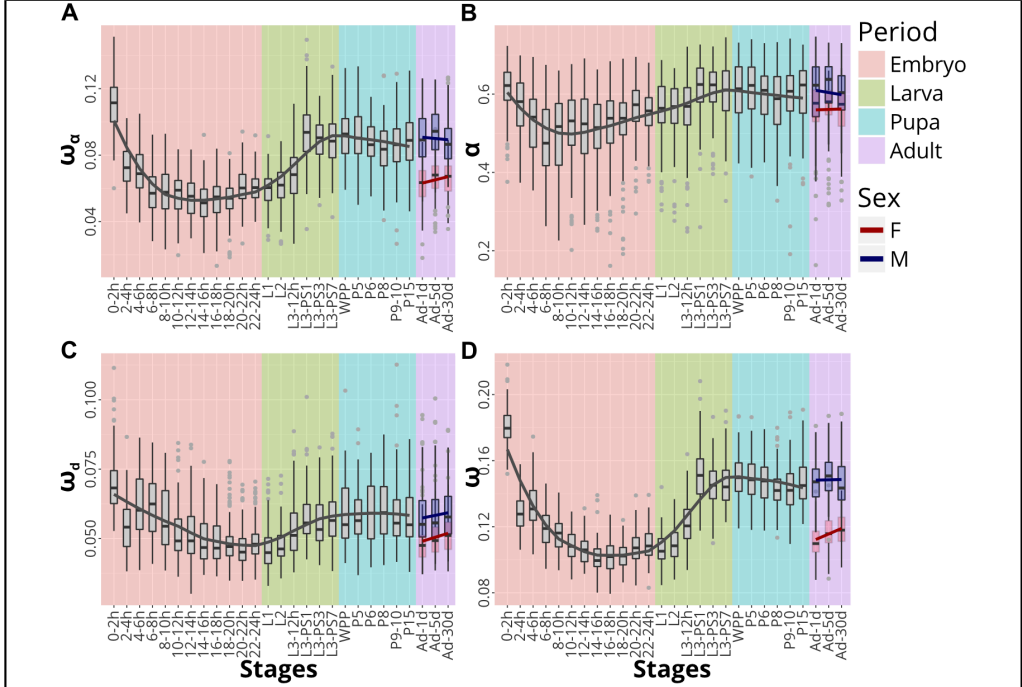


Figure 4.7: ω_α (A), α (B), ω_d (C), ω (D) through the life cycle of *D. melanogaster*. Each time point represents 1,000 random samples of 350 genes (with replacement) expressed in a stage. Red line represents a LOESS regression. Female and male stages are fitted to a linear regression. There are 12 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 $r^2 = 0.622, p = 2.54 \times 10^{-6}$ with the male stages and $r^2 = 0.7, p = 1.67 \times 10^{-6}$ 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: $r^2 = 0.702, p = 1.66 \times 10^{-6}$; $r^2 = 0.867, p = 6.19 \times 10^{-7}$; $r^2 = 0.781, p = 1.27 \times 10^{-6}$; $r^2 = 0.667, p = 1.85 \times 10^{-6}$, respectively) and females (rank correlations: $r^2 = 0.697, p = 1.68 \times 10^{-6}$; $r^2 = 0.888, p = 4.67 \times 10^{-7}$; $r^2 = 0.813, p = 1.04 \times 10^{-6}$; $r^2 = 0.756, p = 1.44 \times 10^{-6}$, 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 high ω_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.

--

<h2>Concluding Remarks</h2>	
-----------------------------	--

References

- Abzhanov, A., W. P. Kuo, C. Hartmann, B. R. Grant, P. R. Grant, and C. J. Tabin (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., C. Ofria, and T. C. Collier (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., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson (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 L. Wolpert (1965). Cybernetics and development I. Information theory. *Journal of Theoretical Biology* 8(2), 244–257.
- Arbeitman, M. N., E. E. M. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science (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., W. Haerty, and R. S. Singh (2009). Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of *Drosophila*. *BMC biology* 7(1), 42.
- Ballard, W. W. (1981). Morphogenetic Movements and Fate Maps of Vertebrates. *American Zoologist* 21(2), 391–399.
- Bell, G. and A. Mooers (1997). Size and complexity among multicellular organisms. *Biological Journal of the Linnean Society* 60(3), 345–363.
- Betancourt, A. J. and D. C. Presgraves (2002). Linkage limits the power of natural selection in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 99(21), 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.
- Bunn, J. M., D. M. Boyer, Y. Lipman, E. M. St Clair, J. Jernvall, and I. Daubechies (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 V. Hartenstein (1985). *The Embryonic Development of Drosophila melanogaster*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Canning, E. U. and B. Okamura (2003). Biodiversity and Evolution of the Myxozoa. *Advances in Parasitology* 56, 43–131.
- Carroll, S. B., J. K. Grenier, and S. D. Weatherbee (2001). *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. Malden, MA.: Blackwell Science.
- Casari, G., C. Sander, and A. Valencia (1995). A method to predict functional residues in proteins. *Nature structural biology* 2(2), 171–8.
- Civetta, A. and R. S. Singh (1995). High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *Journal of molecular evolution* 41(6), 1085–95.
- Cole, A. G. and I. A. Meinertzhagen (2004). The central nervous system of the ascidian larva: mitotic history of cells forming the neural tube in late embryonic *Ciona intestinalis*. *Developmental 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., A. Erives, A. Di Gregorio, A. Chang, and M. Levine (1997). Dorsoventral patterning of the vertebrate neural tube is conserved in a protochordate. *Development* 124(12), 2335–2344.
- 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*

References

- nisms* 1789(4), 248–249.
- Davis, J. C., O. Brandman, and D. A. Petrov (2005). Protein evolution in the context of *Drosophila* development. *Journal of molecular evolution* 60(6), 774–85.
- Dehal, P., Y. Satou, R. K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D. M. Goodstein, N. Harafuji, K. E. M. Hastings, I. Ho, K. Hotta, W. Huang, T. Kawashima, P. Lemaire, D. Martinez, I. A. Meinertzhagen, S. Nacula, M. Nonaka, N. Putnam, S. Rash, H. Saiga, M. Satake, A. Terry, L. Yamada, H.-G. Wang, S. Awazu, K. Azumi, J. Boore, M. Branno, S. Chin-Bow, R. DeSantis, S. Doyle, P. Francino, D. N. Keys, S. Haga, H. Hayashi, K. Hino, K. S. Imai, K. Inaba, S. Kano, K. Kobayashi, M. Kobayashi, B.-I. Lee, K. W. Makabe, C. Manohar, G. Matassi, M. Medina, Y. Mochizuki, S. Mount, T. Morishita, S. Miura, A. Nakayama, S. Nishizaka, H. Nomoto, F. Ohta, K. Oishi, I. Rigoutsos, M. Sano, A. Sasaki, Y. Sasakura, E. Shoguchi, T. Shin-i, A. Spagnuolo, D. Stainier, M. M. Suzuki, O. Tassy, N. Takatori, M. Tokuoka, K. Yagi, F. Yoshizaki, S. Wada, C. Zhang, P. D. Hyatt, F. Larimer, C. Detter, N. Doggett, T. Glavina, T. Hawkins, P. Richardson, S. Lucas, Y. Kohara, M. Levine, N. Satoh, and D. S. Rokhsar (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 M. Levine (1999). Regulation of Ci-tropomyosin-like, a Brachyury target gene in the ascidian, *Ciona intestinalis*. *Development* 126(24), 5599–5609.
- Dobzhansky, T. (1956). What is an Adaptive Trait? *The American Naturalist* 90(855), 337–347.
- Dobzhansky, T. (1968). On Some Fundamental Concepts of Darwinian Biology. In *Evolutionary Biology*, pp. 1–34. Boston, MA: Springer US.
- Domazet-Lošo, T., J. Brajković, and D. Tautz (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 D. Tautz (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468(7325), 815–8.
- Drost, H.-G., A. Gabel, I. Grosse, and M. Quint (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.
- Endler, J. A. (1986). *Natural Selection in the Wild*. Monographs in population biology. Princeton University Press.
- Ernst, J. and Z. Bar-Joseph (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC bioinformatics* 7, 191.
- Eyre-Walker, A. and P. D. Keightley (2007). The distribution of fitness effects of new mutations. *Nature reviews. Genetics* 8(8), 610–8.
- Eyre-Walker, A. and P. D. Keightley (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular biology and evolution* 26(9), 2097–108.
- Eyre-Walker, A., M. Woolfit, and T. Phelps (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173(2), 891–900.
- Ferson, S., F. J. Rohlf, and R. K. Koehn (1985). Measuring Shape Variation of Two-Dimensional Outlines. *Systematic Zoology* 34(1), 59.
- Filipowicz, W., L. Jaskiewicz, F. A. Kolb, and R. S. Pillai (2005). Post-transcriptional gene silencing by siRNAs and miRNAs. *Current Opinion in Structural Biology* 15(3), 331–341.
- Forgacs, G. and S. A. Newman (2005). *Biological Physics of the Developing Embryo*. Cambridge University Press.
- Frise, E., A. S. Hammonds, and S. E. Celniker (2010). Systematic image-driven analysis of the spatial *Drosophila* embryonic expression landscape. *Molecular systems biology* 6, 345.
- Gall, J. G. and M. L. Pardue (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.
- García-Bellido, A. and J. R. Merriam (1969). Cell lineage of the imaginal discs in *Drosophila* gynandromorphs. *Journal of Experimental Zoology* 170(1), 61–75.
- García-Bellido, A., P. Ripoll, and G. Morata (1973). Developmental Compartmentalisation of the Wing Disk of *Drosophila*. *Nature* 245(147), 251–253.
- Gilbert, S. F. (Ed.) (1991). *A Conceptual History of Modern Embryology*. Boston, MA: Springer US.
- 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.*, Dabner 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.). Sinauer Associates.
- Gillespie, J. H. (1994). *The Causes of Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press.
- Gloor, G. B., L. C. Martin, L. M. Wahl, and S. D. Dunn (2005). Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions. *Biochemistry* 44(19), 7156–7165.
- Gould, S. J. (1996). *Full House: The Spread of Excellence From Plato to Darwin*. New York: Harmony Books.
- Graveley, B. R., A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. van Baren, N. Boley, B. W. Booth, J. B. Brown, L. Cherbas, C. a. Davis, A. Dobin, R. Li, W. Lin, J. H. Malone, N. R. Mattiuzzo, D. Miller, D. Sturgill, B. B. Tuch, C. Zaleski, D. Zhang, M. Blanchette, S. Dudoit, B. Eads, R. E. Green, A. Hammonds, L. Jiang, P. Kapranov, L. Langton, N. Perrimon, J. E. Sandler, K. H. Wan, A. Willingham, Y. Zhang, Y. Zou, J. Andrews, P. J. Bickel, S. E. Brenner, M. R. Brent, P. Cherbas, T. R. Gingeras, R. a. Hoskins, T. C. Kaufman, B. Oliver, and S. E. Celniker (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., A. Wilczynska, and J. A. Siedlecki (2001). Regulatory Functions of 3 UTRs. *Biochemical and Biophysical Research Communications* 288(2), 291–295.
- Gunz, P., P. Mitteroecker, and F. L. Bookstein (2005). Semilandmarks in Three Dimensions. In *Modern Morphometrics in Physical Anthropology*, pp. 73–98. New York: Kluwer Academic Publishers-Plenum Publishers.
- Haeckel, E. (1874). *Anthropogenie oder Entwicklungsgeschichte des Menschen*. Engelmann, Leipzig.
- Haeckel, E. (1880). *The history of creation*, Volume 1. New York: Appleton and Company.
- Haerty, W., S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. Ravi Ram, L. K. Sirot, L. Levesque, C. G. Artieri, M. F. Wolfner, A. Civetta, and R. S. Singh (2007). Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. *Genetics* 177(3), 1321–35.
- Hahn, M. W. and G. A. Wray (2002). The g-value paradox. *Evolution and Development* 4(2), 73–75.
- Hammonds, A. S., C. A. Bristow, W. W. Fisher, R. Weiszmann, S. Wu, V. Hartenstein, M. Kellis, B. Yu, E. Frise, and S. E. Celniker (2013). Spatial expression of transcription factors in *Drosophila* embryonic organ development. *Genome biology* 14(12), R140.
- Hannenhalli, S. S. and R. B. Russell (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.
- Hershberg, R. and D. A. Petrov (2008). Selection on codon bias. *Annual review of genetics* 42, 287–99.
- Holder, T. (2010). History of Developmental Biology. In *Encyclopedia of Life Sciences*. Chichester, UK: John Wiley & Sons, Ltd.
- Hotta, K., K. Mitsuhashi, H. Takahashi, K. Inaba, K. Oka, T. Gojobori, and K. Ikeo (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.
- Hudson, C., S. Lotito, and H. Yasuo (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 G. De Beer (1963). *The elements of experimental embryology*. Cambridge comparative 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., K. Hino, K. Yagi, N. Satoh, and Y. Satou (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., M. Levine, N. Satoh, and Y. Satou (2006). Regulatory blueprint for a chordate embryo. *Science (New York, N.Y.)* 312(5777), 1183–7.
- Jacob, F. and J. Monod (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 J. Sharpe (2014). On the concept of

References

- mechanism in development. In A. Minelli and T. Pradeu (Eds.), *Towards a Theory of Development*, pp. 56–78. OUP Oxford.
- James Rohlf, F. and L. F. Marcus (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 L. Sel  ne (1999). Laser confocal microscopy and geographic information systems in the study of dental morphology. *Palaeontologia electronica* 2(1), 18.
- Jolicoeur, P. and J. E. Mosimann (1960). Size and shape variation in the painted turtle. A principal component analysis. *Growth* 24, 339–54.
- Kalinka, A. T. and P. Tomancak (2012). The evolution of early animal embryos: conservation or divergence? *Trends in ecology & evolution* 27(7), 385–93.
- Kalinka, A. T., K. M. Varga, D. T. Gerard, S. Preibisch, D. L. Corcoran, J. Jarrells, U. Ohler, C. M. Bergman, and P. Tomancak (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468(7325), 811–4.
- Kang, T. S. and R. M. Kini (2009). Structural determinants of protein folding. *Cellular and Molecular Life Sciences* 66(14), 2341–2361.
- Kapp, L. D. and J. R. Lorsch (2004). The Molecular Mechanics of Eukaryotic Translation. *Annual Review of Biochemistry* 73(1), 657–704.
- Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937), 241–254.
- Kimchi-Sarfaty, C., J. M. Oh, I.-W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. Gottesman (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 T. Ohta (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 T. H. Jukes (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., T. L. Karr, M. McCutchan, S. J. Newfeld, and S. Kumar (2012). Comparison of embryonic expression within multi-gene 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 C. R. Giardina (1982). Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing* 18(3), 236–258.
- Kumar, S., C. Konikoff, B. Van Emden, C. Busick, K. T. Davis, S. Ji, L.-W. Wu, H. Ramos, T. Brody, S. Panchanathan, J. Ye, T. L. Karr, K. Gerold, M. McCutchan, and S. J. Newfeld (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.
- L  cuyer, E., H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T. R. Hughes, P. Tomancak, and H. M. Krause (2007). Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. *Cell* 131(1), 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., W. C. Smith, and H. Nishida (2008). Ascidians and the plasticity of the chordate developmental program. *Current biology : CB* 18(14), R620–31.
- Levin, M., T. Hashimshony, F. Wagner, and I. Yanai (2012). Developmental Milestones Punctuate Gene Expression in the *Caenorhabditis* Embryo. *Developmental Cell* 22(5), 1101–1108.
- Lichtenstein, F., F. Antoneli, and M. R. S. Briones (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., P.-A. Miquel, C. Sonnenschein, and A. Soto (2012). Is information a proper observ-

- able for biological organization? *Progress in Biophysics and Molecular Biology* 109(3), 108–114.
- Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javadi, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Râmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. A. Gibbs (2012). The *Drosophila melanogaster* Genetic Reference Panel. *Nature* 482(7384), 173–8.
- Mann, M. and O. N. Jensen (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology* 21(3), 255–261.
- Marianes, A. and A. C. Spradling (2013). Physiological and stem cell compartmentalization within the *Drosophila* midgut. *eLife* 2, e00886.
- Martínez-Abadías, N., R. Mateu, M. Niksic, L. Russo, and J. Sharpe (2016). Geometric Morphometrics on Gene Expression Patterns Within Phenotypes: A Case Example on Limb Development. *Systematic biology* 65(2), 194–211.
- Matsuoka, T., T. Ikeda, K. Fujimaki, and Y. Satou (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 M. Kreitman (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351(6328), 652–4.
- 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., F. Serra, N. J. Lavagnino, H. Dopazo, and E. Hasson (2013). Positive selection in nucleoporins challenges constraints on early expressed genes in *Drosophila* development. *Genome biology and evolution* 5(11), 2231–41.
- Messer, P. W. and D. A. Petrov (2013). Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America* 110(21), 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., H. Takahashi, and N. Satoh (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 P. Gunz (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., J. Terai, R. Okubo, K. Hotta, and K. Oka (2012). Three-dimensional anatomy of the *Ciona intestinalis* tailbud embryo at single-cell resolution. *Developmental biology* 372(2), 274–84.
- Nicol, D. and I. A. Meinertzhagen (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., M. L. Wayne, K. L. Harmon, and L. M. McIntyre (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

References

- Substitutions in Evolution. *Nature* 246(5428), 96–98.
- Ohta, T. and J. H. Gillespie (1996). Development of Neutral and Nearly Neutral Theories. *Theoretical Population Biology* 49(2), 128–142.
- Ostrowski, S., H. A. Dierick, and A. Bejsovec (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.
- Peregrín-Alvarez, J. M., C. Sanford, and J. Parkinson (2009). The conservation and evolutionary modularity of metabolism. *Genome biology* 10(6), R63.
- Piasecka, B., P. Lichocki, S. Moretti, S. Bergmann, and M. Robinson-Rechavi (2013). The hourglass and the early conservation models—co-existing patterns of developmental constraints in vertebrates. *PLoS genetics* 9(4), e1003476.
- Plotkin, J. B. and G. Kudla (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics* 12(1), 32–42.
- Poe, S. and M. H. Wake (2004). Quantitative tests of general models for the evolution of development. *The American naturalist* 164(3), 415–22.
- Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchon, J. J. Emerson, P. Saelao, D. J. Begun, and C. H. Langley (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., N. Alhusaini, Y.-H. Chen, S. Martin, N. Morris, N. Kline, S. Olson, D. Weinberg, K. Baker, B. Graveley, and J. Collier (2015). Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* 160(6), 1111–1124.
- Pröschel, M., Z. Zhang, and J. Parsch (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., J. Hanken, M. L. Gooneratne, C. Pieau, a. Raynaud, L. Selwood, and G. M. Wright (1997). There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anatomy and embryology* 196(2), 91–106.
- Roberts, D. B. (1998). *Drosophila: a practical approach*. Practical approach series. IRL Press at Oxford University Press.
- Salazar-Ciudad, I. (2010). Morphological evolution and embryonic developmental diversity in metazoa. *Development (Cambridge, England)* 137(4), 531–9.
- Salvador-Martínez, I. and I. Salazar-Ciudad (2015). How complexity increases in development: An analysis of the spatial-temporal dynamics of 1218 genes in *Drosophila melanogaster*. *Developmental Biology* 405(2), 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., A. Moya, and S. F. Elena (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.
- Sebé-Pedrós, A., A. Ariza-Cosano, M. T. Weirauch, S. Leininger, A. Yang, G. Torruella, M. Adamski, M. Adamska, T. R. Hughes, J. L. Gómez-Skarmeta, and I. Ruiz-Trillo (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., C. N. Cole, M. A. McPeck, and K. J. Peterson (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., O. Binder, and F. L. Conlon (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., P. W. Holland, and C. F. Graham (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.
- Stapleton, M., G. Liao, P. Brokstein, L. Hong, P. Carninci, T. Shiraki, Y. Hayashizaki, M. Champe, J. Pacleb, K. Wan, C. Yu, J. Carlson, R. George, S. Celniker, and G. M. Rubin (2002). The Drosophila Gene Collection: Identification of Putative Full-Length cDNAs for 70% of D. melanogaster Genes. *Genome Research* 12(8), 1294–1300.
- The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* 467(7319), 1061–1073.
- Tokuoka, M., K. S. Imai, Y. Satou, and N. Satoh (2004). Three distinct lineages of mesenchymal cells in Ciona intestinalis embryos demonstrated by specific gene expression. *Developmental biology* 274(1), 211–24.
- Tomancak, P., A. Beaton, R. Weizmann, E. Kwan, S. Shu, S. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. Celniker, and G. Rubin (2002). Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biology* 3(12), 1–14.
- Tomancak, P., B. P. Berman, A. Beaton, R. Weizmann, E. Kwan, V. Hartenstein, S. E. Celniker, and G. M. Rubin (2007). Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome biology* 8(7), R145.
- Valentine, J. W., A. G. Collins, and C. P. Meyer (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., F. Han, G. Wellhagen, B. Dahle, M. Kawata, N. Haddad, Z. L. P. Simões, M. H. Allsopp, I. Kandemir, P. De la Rúa, C. W. Pirk, and M. T. Webster (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.
- Weizmann, R., A. S. Hammonds, and S. E. Celniker (2009a). Determination of gene expression patterns using high-throughput RNA in situ hybridization to whole-mount Drosophila embryos. *Nature protocols* 4(5), 605–18.
- Weizmann, R., A. S. Hammonds, and S. E. Celniker (2009b). Determination of gene expression patterns using high-throughput RNA in situ hybridization to whole-mount Drosophila embryos. *Nature Protocols* 4(5), 605–618.
- Winchester, J. M. (2016). MorphoTester: An Open Source Application for Morphological Topographic Analysis. *PloS one* 11(2), e0147649.
- Winchester, J. M., D. M. Boyer, E. M. St. Clair, A. D. Gosselin-Ildari, S. B. Cooke, and J. A. Ledogar (2014). Dental topography of platyrrhines and prosimians: Convergence and contrasts. *American Journal of Physical Anthropology* 153(1), 29–44.
- Wray, G. A. (2000). The evolution of embryonic patterning mechanisms in animals. *Seminars in cell & developmental biology* 11(6), 385–93.
- Yasuo, H. and N. Satoh (1998). Conservation of the developmental role of Brachyury in notochord formation in a urochordate, the ascidian *Balcyntia 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., D. L. Swiderski, and H. D. Sheets (2012). *Geometric Morphometrics for Biologists: A Primer*. Academic Press. Elsevier Academic Press.
- Zelditch, M. L., A. R. Wood, R. M. Bonett, and D. L. Swiderski (2008). Modularity of the rodent mandible: Integrating bones, muscles, and teeth. *Evolution & Development* 10(6), 756–768.