

**Identification of the molecular changes
underlying head morphology variation in
closely related *Drosophila* species**

Dissertation

For the award of the degree

Doctor rerum naturalium (Dr. rer. nat.)

Division of Mathematics and Natural Sciences

of the Georg-August-Universität Göttingen

within the doctoral program *Genes and Development*

of the Georg-August University School of Science (GAUSS)

submitted by

Montserrat Torres Oliva

from Barcelona (Spain)

Göttingen, March 2016

Thesis Committee:

Dr. Nico Posnien (1st Reviewer, advisor)

Department of Developmental Biology, Johann-Friedrich-Blumenbach-Institute of Zoology and Anthropology, Georg-August-University Göttingen

Prof. Dr. Martin Göpfert (2nd Reviewer)

Department of Cellular Neurobiology, Schwann-Schleiden Research Centre, Georg-August-University Göttingen

Prof. Dr. Tim Beißbarth

Department of Medical Statistics, University Medical Center Göttingen

Further members of the Examination Board:

Prof. Dr. Gregor Bucher

Department of Evolutionary Developmental Genetics, Johann-Friedrich-Blumenbach-Institute of Zoology and Anthropology, Georg-August-University Göttingen

Prof. Dr. Daniel Jackson

Courant Research Centre Geobiology, Georg-August-University of Göttingen

PD Dr. Halyna Shcherbata

Research Group Gene Expression and Signaling, Max Planck Institute for Biophysical Chemistry

Date of oral examination: 23rd May 2016

Declaration

Herewith I declare, that I prepared the Dissertation

“Identification of the molecular changes underlying head morphology variation in closely related *Drosophila* species”

on my own and with no other sources and aids than quoted.

Göttingen, 31.03.2016

Montserrat Torres Oliva

I want to dedicate this work

to my friend Orla Lamlor

Acknowledgements

I would like to take this chance to thank all the people who, in some way or another, I feel have contributed and helped me complete this thesis.

First and foremost, I want to thank my supervisor Dr. Nico Posnien. Thank you for your constant support, encouragement and guidance; for so many interesting and productive discussions, and for always listening and valuing my opinions. Thank you for all what you have taught me and for always finding time to answer my questions. I am very happy I could be the first of the many Ph.D. students that will follow. Vielen, vielen Dank!

I also want to thank the other members of my Thesis Committee, Prof. Dr. Martin Göpfert and Prof. Dr. Tim Beißbarth, for the useful comments and directions for my project during the Thesis Committee Meetings. I am also thankful to Prof. Dr. Gregor Bucher, Prof. Dr. Daniel Jackson and PD. Dr. Halyna Shcherbata for accepting to be part of my extended Committee.

I am grateful to Prof. Dr. Ernst Wimmer for letting me be part of the Department of Developmental Biology. I also want to thank him and Prof. Dr. Gregor Bucher, Prof. Dr. Sigrid Hoyer-Fender, Dr. Nikola-Michael Prpic-Schäper and Dr. Gerd Vorbrüggen for the interesting discussions and constructive input during my Progress Reports. I want to thank especially Dr. Marita Büscher for all her help and advice.

I have been lucky to be able to supervise great students that have also contributed to this work. I want to thank Gordon Wiegleb, Julia Schneider and Felix Kaufholz for their work on the Hunchback project and Elisa Buchberger and Melissa Jüds for the endless hours dissecting discs.

I also want to thank all the members of Lab2 for creating such a nice working environment. Thanks to Natascha for our long scientific (and not so scientific) conversations and for always seeing through my seriousness. I am also very thankful to Christoph, Felix, Elisa, Yan Li, Natalia and Kefei for all their help and friendship. Thanks also to Reya for her visits that always cheered me up.

I would like to thank all the many people that I have met in the Department of Developmental Biology during the last years, including all students, technicians and secretaries. I want to thank especially Ingrid, Sabrina, Alice, Georg, Kolja, Beni, Stefan, Elke and Beate for their help in the lab and for the time we spent together. I also want to thank my friend Anna Stief for introducing me to this Department.

A big Thank You goes to everybody in the group of Dr. Alistair P. McGregor in Oxford for the great work together. Especial thanks to Dr. Isabel Almudi for all you have taught me and for always answering my many questions in record time. I also want to thank my former supervisors Prof. Julio Rozas and Prof. Michael Akam for all they taught me and for their trust and encouragement at the very start of my career.

I want to thank Gabriela Salinas and everybody at the Transcriptome and Genome Analysis Laboratory (TAL) for the useful discussions and for generating all the sequencing data for this project. I also want to acknowledge the people at the GWDG for the great resources they provide.

Somehow, I would also like to thank the very many flies that have unwillingly participated in this work. I thank them also for (almost) always laying eggs when I needed them. This would have been impossible without them.

I am extremely thankful to Fundació Obra Social La Caixa and the Deutscher Akademischer Austauschdienst (DAAD) for granting me the scholarship that allowed me to carry out this work. I was very fortunate as well to receive funding from the Graduate School for Neurosciences, Biophysics and Molecular Biosciences (GGNB), which allowed me to extend my work for three months. I am also thankful to everybody in the GGNB for all their assistance during my thesis.

In my three and a half years in Göttingen I have met some great people and they all made my time here better. I want to especially thank Madlen, Irene and Hans for being the best housemates one could ever wish for.

El que m'ha donat més forces els últims anys han sigut les meves escapades a casa. Vull donar les gràcies a l'Angela, la Clara, la Joana, la Mònica, la Gemma, el Pablo i la Irene perquè en cada una de les meves visites heu trobat temps per veure'ns. Gràcies per fer-me sentir sempre com si ens haguéssim vist ahir.

Als meus pares els hi vull agrair tot. Gràcies per haver-me permès i ajudat a estudiar i per haver-me ensenyat la importància de la ciència i la cultura. Gràcies també pels valors que sempre heu demostrat i que, espero, ara també són els meus. A vosaltres i a la meua germana Irene, gràcies per fer-me riure, per les vostres visites, per les sobretauls i per tot el que hem fet junts. Gràcies també als meus tiets, tietes i cosins per tot el suport i per les millors festes de Cap d'Any. Danke auch an die Familie Hattendorff, dass sie mich so gut aufgenommen haben und für die schönen Tage an der Ostseeküste.

Der größte Dank ist für mein Freund Tobi. Thank you for your love and support and for always making me be the best version of myself. Sense tu, ich hätte es nicht geschafft.

Table of Contents

List of Figures	xii
List of Tables	xv
1 Summary	1
2 Introduction	3
2.1 General introduction	3
2.1.1 Evo-Devo and the study of morphological evolution	3
2.1.2 Gene expression divergence and transcriptomics.....	4
2.1.3 Morphological diversity in insects	5
2.1.4 The model species <i>Drosophila melanogaster</i>	5
2.1.5 <i>Drosophila</i> head structures develop from eye-antennal imaginal discs.....	6
2.1.6 Thesis overview and organization	7
2.2 New regulatory interactions governing <i>Drosophila</i> head development	10
2.2.1 <i>Drosophila</i> head and eye development.....	10
2.2.2 Discovery of new GRN interactions by developmental transcriptomics	12
2.3 A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species	14
2.4 Gene expression divergence in closely related <i>Drosophila</i> species	16
2.4.1 Gene expression divergence, GRN evolution and micro-evo-devo.....	17
2.4.2 Allele-specific expression studies.....	18
2.4.3 Regulatory divergence in developing tissues of three closely related <i>Drosophila</i> species.....	20
2.5 Eye size variation in two closely related <i>Drosophila</i> species.....	22
2.5.1 Eye size variation between <i>D. mauritiana</i> and <i>D. simulans</i>	22
2.5.2 Differences in ommatidia structure.....	23
2.5.3 A quantitative trait locus (QTL) correlates with eye size variation	25
3 Materials and Methods	27
3.1 Fly strains, culture and crosses	27
3.2 Immunohistochemistry	28
3.3 Blood-eye barrier assay.....	29
3.4 <i>In situ</i> hybridization	29
3.5 Optical sectioning of <i>Drosophila</i> heads	31
3.6 RNA-seq and bioinformatics analysis	32
3.6.2 New regulatory interactions governing <i>Drosophila</i> head development.....	34
3.6.3 Gene expression divergence in closely related <i>Drosophila</i> species.....	36
3.6.4 Eye size variation in two closely related <i>Drosophila</i> species	41

4	Results	43
4.1	New regulatory interactions governing <i>Drosophila</i> head development	43
4.1.1	Differentially expressed genes during head development	43
4.1.2	Co-expressed genes during eye-antennal imaginal disc development.....	45
4.1.3	Transcription factors regulating <i>Drosophila</i> head development.....	50
4.1.4	Validation of identified transcription factors.....	52
4.1.5	<i>hb</i> is expressed in retinal sub-perineural glia cells	54
4.1.6	Hb function in the development of retinal glia.....	56
4.1.7	Expression of putative Hb target genes in the eye-antennal imaginal disc.....	63
4.2	A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species	65
4.2.1	Abstract.....	66
4.2.2	Background	67
4.2.3	Results and Discussion.....	70
4.2.4	Conclusions	87
4.2.5	Materials and Methods	88
4.3	Gene expression divergence in closely related <i>Drosophila</i> species	95
4.3.1	Developmental transcriptome of three closely related <i>Drosophila</i> species.....	95
4.3.2	Evolution of gene expression differences	101
4.3.3	Detection of <i>cis</i> and <i>trans</i> regulatory divergence by allele-specific expression (ASE) analysis.....	103
4.4	Eye size variation in two closely related <i>Drosophila</i> species.....	113
4.4.1	Genes differentially expressed between species	113
4.4.2	Expression and functional analysis of candidate genes.....	115
4.4.3	Coding sequence divergence	120
4.4.4	Optical sections of <i>Drosophila</i> heads	120
5	Discussion	123
5.1	New regulatory interactions governing <i>Drosophila</i> head development	123
5.1.1	Dynamic gene co-expression describes eye-antennal imaginal disc developmental events 123	
5.1.2	Enriched <i>cis</i> -regulatory elements in co-expressed genes identify upstream transcription factors.....	126
5.1.3	Description of a new role of Hb in retinal glia development	128
5.1.4	Conclusions and outlook	137
5.2	A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species	139
5.3	Gene expression divergence in closely related <i>Drosophila</i> species	141
5.3.1	Differential gene expression in closely related species.....	141
5.3.2	Expression divergence in developing tissues could be mainly regulated in <i>trans</i>	143
5.3.3	Technical consideration.....	148

5.3.4	Conclusions and outlook	151
5.4	Eye size variation in two closely related <i>Drosophila</i> species.....	154
5.4.1	Identification of candidate genes to regulate eye size differences between closely related species.....	154
5.4.2	<i>ocelliless</i> is the main candidate underlying ommatidia size variation.....	155
5.4.3	Ommatidia structure in <i>D. simulans</i> and <i>D. mauritiana</i>	158
5.4.4	Outlook.....	160
6	References	163
7	Appendix	191
7.1	Abbreviations.....	191
7.2	Supplementary Figures	192
7.3	Supplementary Tables	202
7.4	Sequences of cloned QTL candidates.....	208
8	Curriculum vitae	211

List of Figures

- Figure 2.1.** Adult head structures develop from larval eye-antennal imaginal discs
- Figure 2.2.** Conditions that have been sequenced in this study
- Figure 2.3.** Regulatory types that can give rise to expression divergence
- Figure 2.4.** Eye and face size variation between *D. mauritiana* and *D. simulans*
- Figure 3.1.** Mapping of parental and hybrid reads
- Figure 4.1.1.** Multidimensional scaling plot of *D. melanogaster* samples
- Figure 4.1.2.** Biological Process GO terms enrichment
- Figure 4.1.3.** *D. melanogaster* expression clusters
- Figure 4.1.4.** Networks of genetic interactions
- Figure 4.1.5.** *hb* is expressed in the eye disc in two cells at the base of the optic stalk
- Figure 4.1.6.** *hb* is expressed in the posterior margin of the eye disc
- Figure 4.1.7.** Schematic representation of the carpet glia cells on the eye imaginal disc
- Figure 4.1.8.** *hb* is expressed in sub-perineural glia cells
- Figure 4.1.9.** Cells expressing *hb* migrate through the optic stalk into the disc during larval stages
- Figure 4.1.10.** *hb* and *repo* expression in the brain
- Figure 4.1.11.** *hb* loss of function results in loss of carpet cell nuclei
- Figure 4.1.12.** *hb* loss of function affects axon projection and the organization of other retinal glia cells
- Figure 4.1.13.** *hb* overexpression in wrapping glia
- Figure 4.1.14.** Blood-eye barrier integrity
- Figure 4.1.15.** Hb target genes
- Figure 4.1.16.** Expression of Hb target genes in eye-antennal imaginal discs
- Figure 4.2.1.** Pair-wise length difference between orthologous genes

Figure 4.2.2. Length differences between orthologous genes introduce gene expression biases

Figure 4.2.3. Schematic representation of length bias in inter-species differential expression analysis and our reciprocal re-annotation strategy to correct it

Figure 4.2.4. qPCR results

Figure 4.2.5. Pipeline of reciprocal transcriptome re-annotation method

Figure 4.3.1. Multidimensional scaling plot of three species' samples

Figure 4.3.2. Co-expression clusters in three *Drosophila* species

Figure 4.3.3. Pair-wise differential inter-species gene expression

Figure 4.3.4. Heat map of expression differences between *Drosophila* species

Figure 4.3.5. Differentially expressed genes in the genetic interaction networks

Figure 4.3.6. Mismatches between species references

Figure 4.3.7. Allele-specific expression of mitochondrial genes in the hybrids

Figure 4.3.8. Regulation type

Figure 4.3.9. Overlap of regulation types between eye and wing tissue in *D. melanogaster* x *D. mauritiana* hybrids (96h AEL)

Figure 4.4.1. MA plot of differential gene expression analysis

Figure 4.4.2. QTL region

Figure 4.4.3. Head optical sections

Figure 5.1. Changes in tissue-specific *cis*-regulatory elements of upstream transcription factors are more likely to produce gene expression divergence

List of Tables

Table 3.1. Primer sequences used to clone candidates from *D. simulans*

Table 3.2. Summary of RNA-seq samples

Table 4.1.1. Differentially expressed genes

Table 4.1.2. GO terms of predicted clusters and genetic interactions

Table 4.1.3. i-cisTarget results for each cluster

Table 4.2.1. Number of genes obtained by each annotation method

Table 4.2.2. Differentially expressed genes and correlation between calculated log₂-fold changes and length difference between orthologous genes

Table 4.2.3. Analysis of differential expression

Table 4.2.4. List of RNA-seq samples and the percentage and number of mapped reads to different reference transcriptomes

Table 4.3.1. Mismatches (mm) between orthologs in different references

Table 4.3.2. Mapping stats

Table 4.3.3. GO Terms and transcription factor enrichment of *cis* and *trans* genes between *D. mauritiana* and *D. melanogaster* (96h AEL)

Table 4.4.1. Summary of candidate genes

1 Summary

The great diversity of adult morphologies that we can observe in nature is the product of millions of years of evolution of the underlying developmental programs. The genes that code for the transcription factors and signaling molecules that govern these processes are remarkably conserved across great phylogenetic distances. Thus, it is thought that gene expression divergence is the main driver of morphological evolution. The possibility to study genome-wide patterns of gene expression based on high-throughput transcriptome sequencing (RNA-seq) can provide unprecedented new insights into how the mechanisms that regulate gene expression have evolved to give rise to such outstanding variety in phenotypes.

Insects show a striking morphological diversity, especially in the size and shape of their head and eyes. To understand what parts of the gene regulatory networks that govern head and eye development can evolve to generate morphological differences without disturbing the fundamental developmental programs, a deeper knowledge of these networks is necessary. In the fruit fly *Drosophila melanogaster*, many transcription factors that govern compound eye development are known. However, few target genes of these regulators have been identified, and still little is known about the development of the other organs and cell types that are also part of the fly head. Here I have performed developmental transcriptomics on three key stages of *D. melanogaster* head development in order to obtain a more detailed description of these processes and all the implicated genes. Most interestingly, by gene co-expression analyses I found that the well-known transcription factor Hunchback may play a central role during late eye-antennal imaginal disc development. And indeed, subsequent functional analyses revealed a critical role of Hunchback in the development of a subtype of retinal glia cells that is involved in axon guidance and the formation of an intact blood-brain barrier. This finding and the additional identification of other transcription factors and target genes that I could validate, certify that genome-wide developmental gene co-expression analysis is a powerful tool to increase our knowledge on gene regulatory networks governing developmental processes.

Recent studies have identified significant differences in the size of the heads and compound eyes in the three closely related *Drosophila* species *D. melanogaster*, *D. simulans* and *D. mauritiana*. *D. melanogaster* has a wider face and smaller eyes than its sibling species, while *D. mauritiana* has the biggest eyes and a much narrower face. Therefore, these three species

represent a good model to identify the nodes of the developmental networks that present divergent expression levels that could give rise to adult morphological differences.

Although genomic references are available for these species, the comparability of these resources varied greatly. In order to perform an unbiased inter-species analysis of differential gene expression, I first developed a pipeline to reciprocally re-annotate their genomes. A rigorous benchmarking of this new pipeline in comparison to previously available methods showed that my strategy increased the number of genes that I could compare and it resulted in the most unbiased results. Additionally, this analysis represents the first comprehensive evaluation of existing statistical methods in the context of inter-specific expression divergence.

The unbiased references allowed me to reliably perform a comprehensive transcriptomics analysis to identify all differentially expressed genes between *D. melanogaster*, *D. mauritiana* and *D. simulans* during key stages of head and eye development. By studying allele-specific expression of the viable F₁ hybrids, I could identify the regulatory mechanisms underlying the divergent gene expression between these species. Interestingly, I have found that most gene expression differences in developing tissues are due to changes in the upstream regulatory genes, what is known as variation in *trans*. These results are different to what has been previously reported in adult *Drosophila* tissues and could indicate that different stages of an organism's life are subject to different evolutionary mechanisms influencing gene expression divergence.

Finally, it has been shown that the compound eyes of *D. mauritiana* are bigger than *D. simulans* eyes due to differences in facet size. I have combined available quantitative trait loci data with my genome-wide differential gene expression data to identify the genetic basis of these observed morphological differences. This unbiased strategy in combination with functional tests in *D. melanogaster* has led to the identification of a single gene, namely *ocelliless*, as being the most likely candidate for its regulatory region to have evolved to give rise to the observed morphological differences in eye size.

In conclusion, I could identify new regulatory interactions underlying *Drosophila* head formation. Additionally, I revealed some of the potential molecular changes that may have given rise to morphological diversity. All in all, this work shows how comprehensive transcriptomics analyses can greatly contribute to a better understanding of both developmental and evolutionary processes.

2 Introduction

2.1 General introduction

2.1.1 Evo-Devo and the study of morphological evolution

The great diversity we can observe in all organisms that live on Earth is the result of millions of years of evolution acting on the development of different body plans and morphologies. To understand how the different phenotypes that are present in nature have appeared is one of the main objectives of evolutionary studies. However, in order to understand what the underlying molecular basis of these changes is, a deeper knowledge of the developmental processes that lead to the final phenotypes is required. Evolutionary developmental (evo-devo) studies have been extensively used to understand how the evolution of different genotypes gives rise to different morphologies through changes in developmental processes (Gould, 1977; Raff and Kaufman, 1983). One of the most important findings is that, despite the impressive variety of morphologies that can be observed in nature, a relatively small set of highly conserved genes is responsible to regulate most of the developmental events that give rise to the different body plans (Wagner, 2007). This set of genes is known as the genetic “toolkit”, and it involves mainly transcription factors and signaling pathways (Carroll, 2001). The coding sequence of most of these genes is incredibly well conserved across the metazoan phylogeny. This is shown by the fact that the orthologs of many of these transcription factors can be exchanged between very distantly related species and they can still correctly perform most of their functions (Grens et al., 1995; Halder et al., 1995; Malicki et al., 1990; McGinnis et al., 1990). But if the genetic structure of the main orchestrators of development is so well conserved, how could the current striking morphological diversity evolve? A large body of evidence indicates that the main source of morphological variation comes from differences in how these “toolkit” genes are regulated. That is, morphological diversity arises by divergence in the non-coding regions of genes to change their expression domains in terms of time, place or expression levels (Britten and Davidson, 1971; Carroll, 1995, 2008; King and Wilson, 1975; Prud’homme et al., 2007).

2.1.2 Gene expression divergence and transcriptomics

Understanding the genetic basis of gene expression divergence is a challenging task because it can be regulated at very different levels. On the one hand, it can be caused by variation at the locus of the gene that shows expression divergence, i.e. by changes in its *cis* regulatory region that affect the binding of the transcription factors that activate, repress or enhance its expression at a specific time and place. On the other hand, it can be caused by changes in the upstream gene that regulates its expression, what is called a change in *trans*, since the underlying molecular change that causes this divergence can be in any location on the genome, also far away from the gene locus. Additionally, changes in *trans* can be caused both by changes in the coding region of the upstream transcription factor or by changes in the regulation of this transcription factor (which would be also changes in *cis*). It is a long-standing question whether morphological evolution is more often caused by *cis* or *trans* changes (Wittkopp et al., 2004), and examples of both types of regulation causing gene expression divergence and morphological diversity have been described (e.g. Belting et al., 1998 for *cis* and Löhr and Pick, 2005 for *trans*).

There are many different methods to study gene expression. Traditionally, methods like Northern Blot (Alwine et al., 1977) and *in situ* hybridization (Gall and Pardue, 1969) have been used to detect gene expression, as well as quantitative real-time PCR methods (Bustin, 2000). These methodologies can be used to study the expression of specific genes of interest, but cannot be used for genome-wide analyses. The development of microarray technology allowed analyses of gene expression of thousands of genes at the same time, provided that one synthesizes the corresponding sequences and creates a chip to hybridize them onto (Fan et al., 2004). New technological advances allowed the development of what is known as “second-generation sequencing” (Margulies et al., 2005), a name used to distinguish it from Sanger sequencing, the “first-generation sequencing” (Sanger et al., 1977). These methods are based on sequence amplification and high-throughput sequencing. High-throughput sequencing of *in vitro* transcribed RNA (RNA-seq) is one of the main applications of this technology (Nagalakshmi et al., 2008; Wang et al., 2009). RNA-seq can provide a snapshot of all transcripts present at a specific stage, tissue or cell type and genotypic condition, and in the recent years it has become widely used, mainly due to its dropping costs (Wetterstrand, 2016). Most interestingly, RNA-seq can be used on any organism, provided that a genomic or transcriptomic reference is available, to interrogate the expression of its complete set of transcripts, regardless of previous biological knowledge on that species. This makes this technology the tool of choice for groups

working, for example, with non-model organisms, and makes it especially useful for evolutionary studies (Brawand et al., 2011; Hornett and Wheat, 2012; McManus et al., 2010).

2.1.3 Morphological diversity in insects

Insects are the most diverse animal group and more than half of all the described species of organisms belong to this group. Their body is divided in three parts: the head (a fusion of several segments), three thoracic segments, which harbor three pairs of legs, and the abdomen. Apart from this conserved body structure, insects show an incredible diversity of morphologies, for example, the presence or absence of wings or horns, very different pigmentation patterns or highly specific mouth parts, which represent adaptations to different feeding behaviors (Chapman, 1998; Snodgrass, 1935). This incredible diversity and plasticity has allowed them to adapt to almost all possible environments on Earth. A stunning diversity of head and eye shape can also be observed among insect species. For instance, a case of directional evolution can be observed in male flies of the genus *Zygothrica* (Drosophilidae), where the width of the cuticle between their eyes (subsequently called face) and the angle in which their eyes are oriented gradually increases with taxonomical distance (Grimaldi, 1987). All insects have compound eyes, which are constituted by multiple small subunits called ommatidia. The number of ommatidia per eye can range from fewer than 6 in some worker ants to more than 25,000 in dragon fly species. Even between closely related species or sexes of the same species this number can vary (Posnien et al., 2012; Talarico et al., 2011). These examples of diversity have long fascinated scientists, who have been studying these organisms for centuries.

2.1.4 The model species *Drosophila melanogaster*

The fruit fly *Drosophila melanogaster* is the most extensively studied insect species by far. A great amount of the knowledge we have of arthropod, invertebrate or even metazoan physiology and development comes from studies on this model species and a large percentage of the genes of this species have been studied. *D. melanogaster* was also one of the first species to have its genome sequenced (Adams et al., 2000). Currently, *D. melanogaster* has one of the best quality assembled genome and genome annotation, and both are regularly being updated by the FlyBase Consortium (St. Pierre et al., 2014; dos Santos et al., 2014). FlyBase houses also a well curated website with all current knowledge on this and some of its closely related species. Some years ago also the modENCODE

project launched to facilitate and promote various genome-wide analyses to contribute to a better understanding of genome organization and regulation (Celniker et al., 2009). Finally, the genomes of other *Drosophila* species are also available. The initial project of sequencing 12 different *Drosophila* species, ranging from *D. melanogaster* to *D. grimshawi* (which diverged 40 million years ago) was later followed by the sequencing of other fruit fly species, like *D. mauritiana* (Nolte et al., 2013) or *D. americana* (Fonseca et al., 2013), among many others. All these resources make *D. melanogaster* and its related species one of the most useful model species for all kinds of biological research, including developmental and evolutionary studies.

2.1.5 *Drosophila* head structures develop from eye-antennal imaginal discs

Drosophila are Dipteran species, and as such they are holometabolous insects, meaning that they undergo complete metamorphosis during development to change from the larval into the final adult morphology. Most of the epidermal adult structures of *Drosophila* develop from imaginal discs, which are sac-like tissues that grow during larva and pupa stages and evert during metamorphosis to give rise to the adult organs, such as legs, wings, genitalia or the head and eyes (Fristrom and Fristrom, 1993). The imaginal discs are formed by two layers: the disc proper or imaginal epithelium, where the main patterning and differentiation processes take place, and a squamous layer called peripodial epithelium, which during metamorphosis participates in the eversion and fusion of the imaginal discs (Fristrom and Fristrom, 1975).

The eye-antennal imaginal discs of *Drosophila* give rise to the different head structures, including the eyes, ocelli, antennae and maxillary palps (reviewed in Haynie and Bryant, 1986) (Figure 2.1). These discs have served for extensive research on primordia fate determination, since the initially uniform disc gives rise to structures that are functionally completely different (such as the head capsule, the eyes and the antenna) and all of them develop in the same tissue, where they differentiate and grow adjacent to one another. Regional specification is achieved by the interplay of different gene regulatory networks (GRNs) that generally promote a specific fate (for instance retinal fate), while repressing another (for instance head capsule or antennal fate) (Weasner and Kumar, 2013). Many different GRNs are involved in this process to control differentiation, proliferation and growth, for instance the Notch pathway (Cho and Choi, 1998), the EGFR pathway

(Freeman, 1994) and cell cycle genes (Lopes and Casares, 2015) or the complex network of retinal determination genes (reviewed in Kumar, 2009; Treisman, 2013; see below).

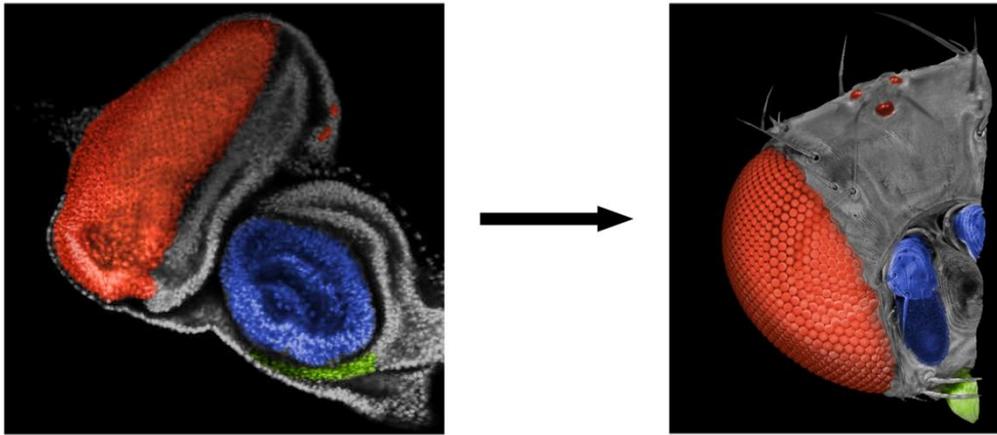


Figure 2.1. Adult head structures develop from larval eye-antennal imaginal discs. Eye-antennal imaginal discs (picture on the left) develop during larval stages (from 22h to 120h after egg laying) and pupal stages to give rise to the adult head (picture on the right). False-color schematic represents the correspondence of the different organ primordia with the adult organs that they will develop into. In red the compound eye and ocelli, in blue the antenna, in green the maxillary palp and in grey the head capsule.

2.1.6 Thesis overview and organization

In order to better understand and describe the processes that take place during *Drosophila* eye-antennal imaginal disc development, I sequenced the transcriptome of three relevant stages: late LII stage (72h after egg laying (AEL)), when the early patterning of the disc finishes; mid LIII stage (96h AEL), at the middle of the process of photoreceptor differentiation; and late LIII stage (120h AEL), at the end of photoreceptor differentiation (Figure 2.2). The comparison of the expressed transcripts at each of these stages can provide a better insight into all the relevant events and key regulators of this process and can also shed light on new regulatory interactions.

Although the GRNs that control head and eye development in *Drosophila* have to be tightly controlled to ensure proper functionality of all organs, they must also be flexible enough to allow the variation that has given rise to the different morphologies that can be observed in adult fly heads and eyes. Therefore, I have also sequenced the eye-antennal imaginal discs' transcriptomes of two closely related species, *Drosophila simulans* and *Drosophila mauritiana* at the same developmental stages (Figure 2.2). These species diverged from *D. melanogaster* less than 3 million years ago, but significant differences in the size of their eyes and in the width of their face have been described (Arif et al., 2013; Hilbrant et al., 2014; Posnien et al.,

2012). A comprehensive transcriptomics analysis of this complex dataset was used to identify the sets of conserved genes and thereupon the flexible nodes of the underlying GRNs that govern head and eye development.

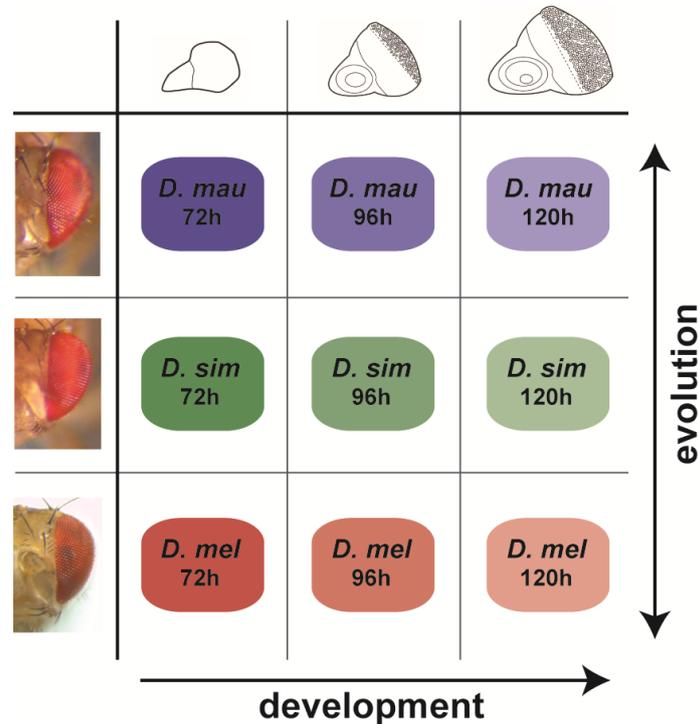


Figure 2.2. Conditions that have been sequenced in this study. The transcriptomes of eye-antennal imaginal discs of three developmental times (late LII, 72h; mid LIII, 96h; late LIII, 120h) and three species (*D. mauritiana*, *D. simulans* and *D. melanogaster*) have been sequenced. The comparison of the transcriptomes across the three stages can provide information on the developmental processes taking place in this tissue (arrow “development”). The comparison of the transcriptomes across the three species can identify the core of genes with conserved gene expression and the variable nodes that allow morphological variation (arrow “evolution”).

This thesis comprises four projects where I have used different approaches to study the development and/or the evolution of the head and eyes of the fly *D. melanogaster* and its closely related species *D. simulans* and *D. mauritiana*. Each of the sections of this thesis (i.e. Introduction, Materials and Methods, Results and Discussion) is divided in four parts, corresponding to the different projects, and they appear in the same order in all sections.

The first project is entitled “**New regulatory interactions governing *Drosophila* head development**”. It contains the developmental transcriptomics analysis of eye-antennal imaginal discs of *D. melanogaster* and the in-depth analysis of a newly discovered role of the transcription factor Hunchback in the development of retinal glia cells.

The second project, “**A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across**

closely related species”, is the full description of a new method that I have developed to enable the inter-species analyses that were required in the third and fourth projects. This method is shortly introduced in the Introduction section and its implications discussed in the Discussion section. In the Results section (section 4.2.), the original manuscript written by me and my supervisor Dr. Nico Posnien can be found as it has been submitted to *BMC Genomics*, where it is currently under revision (minor revision).

The third project is entitled “**Gene expression divergence in closely related *Drosophila* species**”. It consists of a comprehensive analysis of the evolution of the transcriptomes of the three closely related species *D. melanogaster*, *D. mauritiana* and *D. simulans* during key events of eye-antennal imaginal disc development. Moreover, an allele-specific expression analysis is described, which has provided new insights into the different types of regulatory changes that give rise to expression divergence during early developmental processes among closely related species.

Finally, the fourth project is entitled “**Eye size variation in two closely related *Drosophila* species**”. This describes the analysis performed to reveal the genetic basis of the differences in ommatidia size observed between the closely related species *D. simulans* and *D. mauritiana*.

2.2 New regulatory interactions governing *Drosophila* head development

The most prominent parts of a fly's head are its large compound eyes. The events that define the development of *Drosophila* compound eyes have received much more attention than any other region of the eye-antennal imaginal disc. The study of this organ has produced a lot of our current knowledge on cell differentiation and the development of the visual circuitry (Rister and Desplan, 2011; Sanes and Zipursky, 2010). *D. melanogaster* has around 800 ommatidia in each compound eye, and each ommatidia is composed of 8 photoreceptors (PRs), four cone cells and two primary pigment cells, forming a compact cluster (Waddington and Perry, 1960). Two photoreceptors (R7 and R8) are in the center of the cluster, being R8 below R7, and project their axons to the brain medulla, the area responsible for color vision; the other six photoreceptors (R1-R6) surround R7 and R8 and project their axons to the lamina, which is the brain region responsible for motion detection (Wolff and Ready, 1993). Each photoreceptor forms a rhabdomere in its apical region, which is a tightly folded membrane that harbors Rhodopsins, the protein receptors that detect the light photons (Leonard et al., 1992). The cone cells secrete the lens that is located at the top of the ommatidium and the pigment cells isolate the light that each ommatidium receives. Additional secondary and tertiary pigment cells are shared between adjacent ommatidia and contribute to this isolation (Burnet et al., 1967; Wolff and Ready, 1993).

2.2.1 *Drosophila* head and eye development

Not only the compound eyes develop from the eye-antennal imaginal disc, but also the antenna, the maxillary palps, the ocelli and the head cuticle. During LI (1st larval stage), imaginal disc cells ubiquitously express the “eye selector genes” *eyeless* (*ey*) and *twin of eyeless* (*toy*), which are paralogues of the mammalian *Pax6* gene (Gehring, 2002), and the homeodomain transcription factor *homothorax* (*bth*) (Pai et al., 1998; Rieckhof et al., 1997). In LII stage (2nd larval stage), the expression of *ey* and *toy* gets restricted to the posterior part of the disc, where the eye will later develop, and at the anterior region expression of the gene *cut* is activated, marking the future antenna region (Kenyon et al., 2003). *Cut* and *Ey/Toy* repress each other to pattern the antenna and eye primordia, respectively (Punzo et al., 2004). *Cut* activates expression of *Distalles* (*Dll*) and *bth*, which together promote antennal fate (Casares and Mann, 1998; Dong et al., 2000). In parallel to these events,

during late LII stage, in the eye region of the imaginal disc the expression of “early retinal genes” starts to promote retinal differentiation (Kenyon et al., 2003; Kumar and Moses, 2001).

A critical time point for cell fate decisions in the different organ primordia is between late LII and early LIII (Weasner and Kumar, 2013). While the anterior third of the eye-antennal imaginal disc will give rise to the antenna and maxillary palps, the posterior two thirds of the disc contain the compound eye and the face area, with the ocelli developing at the dorsal margin outside the compound eye field (Figure 2.2). At the start of LIII stage, retinal differentiation starts at the posterior margin of the eye disc. This can be clearly detected by the appearance of a transient indentation on the apical surface of the disc, known as morphogenetic furrow. This furrow moves anteriorly as photoreceptor differentiation progresses and it marks the separation between undifferentiated, proliferating cells (or arrested in G1 directly anterior to the morphogenetic furrow (Wolff and Ready, 1993)) from differentiated clusters of retinal cells posterior to the morphogenetic furrow. The relative sizes of the eye and face are determined during this process, and are mainly regulated by the expression of *wingless* (*wg*). In short, *wg* expression at the dorsal and ventral margins of the central third of the disc acts to repress *decapentaplegic* (*dpp*) (which is expressed at the morphogenetic furrow and promotes its progression) (Royet and Finkelstein, 1996) and at the same time promotes expression of *pannier* (*pnr*) (Maurel-Zaffran and Treisman, 2000), *hedgehog* (*hh*) (Domínguez and Hafen, 1997) and *bth* (Pichaud and Casares, 2000). This expression, thus, represses eye tissue in favor of face tissue (Baonza and Freeman, 2002; Ma and Moses, 1995; Treisman and Rubin, 1995).

The cell fate of each type of photoreceptor (R1-R8) inside each ommatidial cluster is determined by cell-cell interaction mechanisms (Ready et al., 1976). The proneural protein Atonal (Ato) is the one responsible to initially single out the cell that will become R8 from an arranged cluster of undifferentiated cells in the morphogenetic furrow, called “rosettes”. This cell will then step-wise recruit R2 and R5 cells to the cluster, followed by R3 and R4, next R1 and R6 and finally R7 (Wolff and Ready, 1993). This process of cell fate determination by cell contact is regulated by the activation of two pathways, Notch and EGFR (Brennan and Moses, 2000; Freeman, 1997), which contribute to spreading the signaling cascades concentrically in the cluster in order for each developing photoreceptor to activate the correct set of genes. Retinal differentiation ends at the end of LIII stage, before pupariation. At that time morphogenetic furrow progression stops and all photoreceptor cells are already differentiated into the correct cell type (Cagan and Ready,

1989). However, these cells continue to develop during pupal stages, for example to express the specific Rhodopsin receptor proteins to populate the rhabdomeres (Wernet et al., 2006) and also programmed cell death takes places to remove inter-ommatidia cells that will not develop into pigment cells (Cagan and Ready, 1989).

2.2.2 Discovery of new GRN interactions by developmental transcriptomics

Developmental processes involve the interplay of large numbers of different molecules that need to be tightly regulated, as they require for each gene to be expressed at the right time, at the right place and in the correct amount. Transcription factors are the main orchestrators of these processes, as they regulate the correct expression of other genes. Transcription factors bind to enhancer elements of their target genes and in that way they activate or repress their expression (reviewed in Lemon, 2000; Spitz and Furlong, 2012). Enhancer elements are usually bound by more than one transcription factor, and therefore this regulation can be better fine-tuned. Developmental gene regulatory networks (GRNs) represent the interactions between transcription factors, their binding sites and the targets they regulate (Davidson, 2006; Davidson et al., 2002) In the era of high-throughput techniques, interactions between transcription factors and their targets genes can be inferred by gene expression profiling. For instance, reverse genetics strategies are usually used to remove the expression of specific transcription factors and to identify which genes show an effect on their expression levels after this perturbation (Marbach et al., 2012). This analysis can reveal direct and indirect target genes of the studied transcription factors. In order to test whether these interactions are direct (the transcription factor directly binds to the regulatory DNA sequence of the target genes), chromatin immunoprecipitation analysis can be performed with a transcription factor of interest, followed by deep sequencing of the regions this transcription factor binds to (ChIP-seq) (Johnson et al., 2007). This method can unravel direct interactions between transcription factors and their binding sites. However, the described approaches require previous knowledge of the transcription factors that are involved in the developmental process of interest.

As it has been described above, some of the main transcription factors governing *Drosophila* head development are known, especially for the differentiation of compound eye photoreceptors (Domínguez and Casares, 2005; Kumar, 2009; Treisman, 2013). An extensive study to describe the GRN underlying photoreceptor differentiation has been recently published (Potier et al., 2014a). This was based on the analysis of 72 different

transcription factor perturbations and transcriptome sequencing of posterior eye-antennal imaginal disc tissue, which allowed the identification of more than 5,000 direct transcription factor-gene interactions. However, this approach only provided information of the regulatory events taking place in photoreceptor cells, as only the transcriptome of cells expressing photoreceptor specific genes were sequenced (Potier et al., 2014a). Many other cell types are present in the eye-antennal imaginal discs such as undifferentiated, proliferating cells, cells that will give rise to head cuticle or to the mouth parts, antennal precursors, including other types of neurons, and also glia cells that support these neurons (Choi and Benzer, 1994; Haynie and Bryant, 1986; Jurgens and Hartenstein, 1993). Especially, very few genes involved in the important transition from LII stage eye-antennal imaginal discs to LIII stage ones are currently known.

In order to obtain a better understanding of these transitions I have incorporated developmental high-throughput data (i.e. at different consecutive time points) into the current knowledge of the different networks that coordinate *Drosophila* head development. I have performed a comprehensive genome-wide analysis of the expression profiles to identify groups of genes that are dynamically co-expressed across the different stages. Since these modules of co-expressed genes can appear as a result of the action of upstream co-regulators, I combined these data with known information about transcription factor-DNA and transcription factor-gene interactions to identify some of these upstream factors. This developmental transcriptomics analysis has provided a list of putative regulators of *Drosophila* head development, some of which have not been previously described to have a function in this process and therefore I have tested their possible role during eye-antennal imaginal disc development.

2.3 A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species

Since it has been proposed that morphological divergence may be mainly the result of variation in expression of a limited number of highly conserved “toolkit” genes, it is of major interest to study genome-wide expression differences among species (Carroll, 2001, 2008; King and Wilson, 1975). Thus, in evolutionary studies that make use of RNA-seq technology, transcriptomic data from different species is compared. To obtain reliable results in this kind of analyses it is of utmost importance to use unbiased references for each of the sequenced species and it has already been recognized that this can pose a challenge (Musser and Wagner, 2015; Roux et al., 2015), mainly due to the lack of references for some non-model species or due to the different qualities of these references. Inter-species RNA-seq-based analyses of differential gene expression have already been performed, but they have mostly focused on a small set of highly conserved genes or have only analyzed general transcription patterns (Brawand et al., 2011; Busby et al., 2011; McManus et al., 2010; Rifkin et al., 2003). But for an unbiased genome-wide comparison of gene expression profiles, it is important to study gene expression between all orthologous genes of the analyzed species.

My aim was to compare gene expression levels across three closely related *Drosophila* species, *D. melanogaster*, *D. mauritiana* and *D. simulans*. However, I could recognize that the quality of the genome annotation of *D. melanogaster* was of higher quality than the annotation of the other two non-model species. In particular, a large number of annotated genes in *D. mauritiana* and *D. simulans* were truncated, mainly due to assembly errors, and therefore were shorter than their *D. melanogaster* orthologs. A large number of statistical methods have been developed to reliably identify genes that are significantly differentially expressed between two or more conditions of interest based on RNA-seq data (e.g. Chu et al., 2015; Love et al., 2014a; Ritchie et al., 2015; Robinson et al., 2010; Trapnell et al., 2012). In general, these analyses are performed comparing different tissues, different time points or control versus diseased or mutant conditions, and therefore the reference used to map the RNA-seq reads is the same in all compared conditions. To compare the relative expression of genes within one sample, researchers have usually applied RPKM-based (reads per kilobase per million reads) methods, where the number of counts mapped to a gene is divided by the length of that gene. This, in principle, corrects for the fact that

longer genes have more reads that map to them, which does not indicate higher expression level. However, the use of these methods has been discouraged (Dillies et al., 2012). It has been shown that, even after this correction, longer genes appear more frequently as significantly higher expressed than shorter genes (Oshlack and Wakefield, 2009). Additionally, it has not been shown yet whether this correction can or should be used in inter-species analyses of differential gene expression to correct for differences in the length of orthologous genes.

In order to overcome these challenges, I have developed a pipeline to reciprocally re-annotate the genomes of *D. melanogaster*, *D. mauritiana* and *D. simulans*. This project is included in this Thesis as a manuscript which is currently in revision in *BMC Genomics*. Please note that only after the development of this pipeline, the analyses described in the sections “**Gene expression divergence in closely related *Drosophila* species**” and “**Eye size variation in two closely related *Drosophila* species**” could be reliably performed, since they both required the comparison of gene expression between the closely related *Drosophila* species.

2.4 Gene expression divergence in closely related *Drosophila* species

The striking morphological diversity present in animals is the result of millions of years of evolution. Evo-devo studies have demonstrated that many processes and their underlying genes are conserved (developmental “toolkit” of genes) (Carroll, 2001; Halder et al., 1995). However, even if developmental processes have been conserved over large phylogenetic distances, they need to be flexible to allow for the incredible diversity of morphologies that exist in nature. An interesting and recurring question in biology is how can GRNs, which need to be tightly controlled to perform the biological processes that allow an organism to develop and live normally, can also be flexible enough to generate inter-species morphological differences. As previously mentioned, many evo-devo studies have shown that a main driver of speciation, especially to generate morphological differences, is gene expression divergence (Carroll, 1995; King and Wilson, 1975). Traditionally, the study of coding sequence evolution has been preferred, as changes in nucleotide sequences can be directly linked to protein sequence divergence (McGinnis et al., 1984; Quiring et al., 1994; Scott et al., 1989). Especially with the sequencing and assembly of new genomes, these studies are relatively straightforward. In contrast, comparing expression at the transcript or protein level across different species poses more difficulties, for instance due to the difficulties of properly normalizing expression levels across different species (Wolf et al., 2010). It is even more challenging to identify the molecular basis of the detected expression differences, since the genetic code of *cis* regulatory elements, if existing, is still largely unknown (Wray, 2007; Yáñez-Cuna et al., 2013). Some studies have been already performed to compare expression levels between orthologous genes across different species (e.g. in yeast species (Busby et al., 2011), mammalian species (Brawand et al., 2011) or fly species (Suvorov et al., 2013)). Still a common standard on how to best perform this kind of analyses, both for the experimental design and for the required subsequent bioinformatics and statistics analyses, does not exist, and it is often complicated to compare results obtained by different groups. What this type of studies have already revealed is that an almost linear correlation between phylogenetic distance and gene expression divergence exists (Khaitovich et al., 2006). And even between very closely related species extensive differences in expression levels of orthologous genes have been detected (McManus et al., 2010).

2.4.1 Gene expression divergence, GRN evolution and micro-evo-devo

What is interesting, however, is not only to show that differences in gene expression exist among orthologs, but rather what kind of regulatory changes are more likely to give rise to morphological differences and get fixed in the genome of the different species (Stern, D. L. and Orgogozo, 2009; Wray, 2007). In other words, what parts of the GRNs underlying the development of an organism's morphology are more likely to evolve? The relationship between network topology and evolution has been studied at the protein level, and some analyses have been performed to investigate if genes with many connections are less likely to be under positive selection than terminal genes which have fewer connections (Siegal et al., 2007). Studies mostly conclude that there is no clear correlation between gene connectivity and amino acid changes (Davila-Velderrain et al., 2014; Montanucci et al., 2011). However, very few such studies have been performed at the gene expression level, mainly due to the lack of high confidence knowledge on conserved networks available for different species and also due to the previously mentioned difficulties of analyzing inter-species gene expression variation.

One way of tackling the lack of available network information is to analyze the type of regulatory variation that generates expression differences between orthologous genes. That is, to determine whether the underlying cause of a gene's expression difference is a change in its *cis* regulatory sequence or if it is a change somewhere else acting in *trans*. This can tell us if the gene expression changes only for that gene or because another upstream factor has changed, and therefore likely affects other gene's expression as well. Different methods can be used for this kind of studies in a genome-wide manner, such as expression quantitative trait loci (eQTL) mapping (Brem et al., 2002) or genome-wide association studies (GWAS) (Dixon et al., 2007). However, these methods demand great effort to create the required mapping population and the ability for the studied species to give rise to viable, fertile hybrids. Moreover, these methods are used to find a link between gene expression and sequence divergence, but this link relies on a relatively arbitrary measure of distance between the polymorphism and the gene with expression differences (Gibson and Weir, 2005). A method that can more precisely classify the type of regulatory variation between orthologs is the study of allele-specific expression (ASE) in hybrid animals (Cowles et al., 2002; Wittkopp et al., 2004).

The use of distantly related species for evo-devo studies can seem more appealing because usually morphological diversity is more pronounced, and also trait innovations are more common. However, the possibility to identify the underlying cause of this divergence at the

nucleotide level gets reduced by evolutionary time (Erives and Levine, 2004; Richards et al., 2005). Therefore, all analyses aiming at identifying the exact molecular changes underlying morphological diversification can only be performed between closely related species, where crosses among them can still produce viable hybrids (ASE studies (Cowles et al., 2002)) or even fertile offspring (eQTL, GWAS (Erickson et al., 2004; Gibson and Weir, 2005)). Micro-evo-devo can be regarded as the study of within species variation or the study of very closely related species (Johnson, 2007; Nunes et al., 2013). This kind of analyses can provide better insight into how natural selection works at the initial steps of speciation to generate morphological diversity (Filteau et al., 2013).

2.4.2 Allele-specific expression studies

In general, the analysis of allele-specific expression (ASE) consists of the distinction of the relative contribution to gene expression of each of the two alleles of a gene in a diploid cell (Knight, 2004; Yan et al., 2002). This kind of analyses are often used in epigenetic studies, for example to identify alleles that are silenced due to chromatin modifications (e.g. Wedd et al., 2015; Wei and Wang, 2013) or to identify imprinted genes, that is genes whose expression depends on the sex of the parent that has contributed them (e.g. Raissig et al., 2011; Skaar and Jirtle, 2015; Mott et al., 2014). In evolutionary studies, ASE analysis can be used to infer the relative contribution of regulatory changes in *cis* and regulatory changes in *trans* on gene expression divergence (Cowles et al., 2002; Wittkopp et al., 2004; Yan et al., 2002). This is based on the fact that in the F₁ hybrid environment, where no recombination has taken place, each allele is still under control of its *cis* regulatory elements, but the *trans* regulatory environment is the same for the two alleles (Figure 2.3). To classify the type of regulatory change driving expression divergence, the relative expression of the orthologous genes in each wild type species (parents) is compared to the relative expression of each allele in the hybrid individuals (Cowles et al., 2002; McManus et al., 2010; Wittkopp et al., 2004). Thus, if the differential gene expression in the parents is also present for the two alleles in the F₁ hybrid, the expression of this gene is assumed to be divergent due to changes in *cis* (Figure 2.3B). In contrast, if a gene is differentially expressed in the parents but the two alleles have equal expression levels in the hybrid environment, the gene is assumed to have expression divergence due to changes in *trans*, i.e. the change is in one of the upstream factors that control its expression. Other types of regulatory changes can also be distinguished with this method, for example *cis* and *trans* changes are assumed to interact (*cis* x *trans*) when a gene shows differentially higher expression in one species in the parents

but in the hybrid the allele from the other species has higher expression (Figure 2.3C). Finally, compensatory regulation is assumed to take place when the alleles are differentially expressed in the F₁ hybrids but the orthologs have equal expression in the parents (Figure 2.3D).

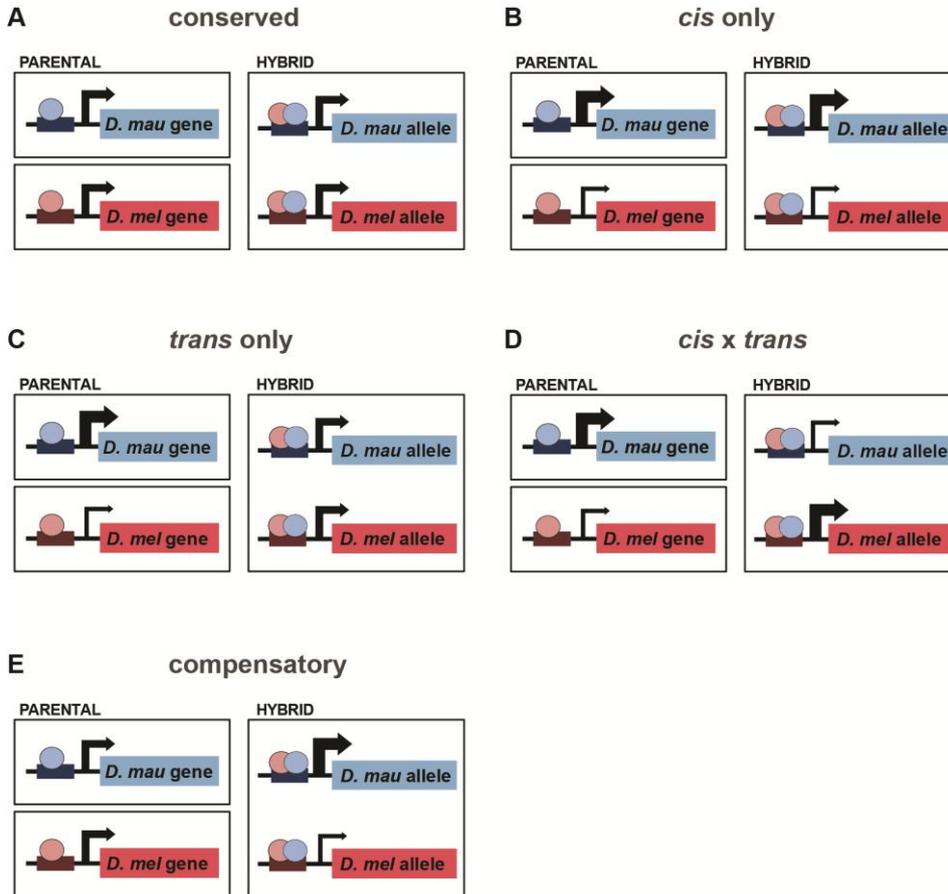


Figure 2.3. Regulatory types that can give rise to expression divergence. Blue circles represent *D. mauritiana* transcription factors and red circles represent *D. melanogaster* transcription factors. Small colored boxes represent the *cis*-regulatory elements that control expression of the downstream gene (large colored boxes). In the hybrid, the transcription factors from both parents can bind to the corresponding *cis*-regulatory elements, but this regulatory region controls the expression of only the corresponding allele. Arrow thickness represents expression level. **(A)** Conserved expression. **(B)** Divergence due to variation in *cis*. **(C)** Divergence due to variation in *trans*. **(D)** Divergence due to *cis* x *trans* variation. **(E)** Compensatory variation. Figure adapted from McManus et al. 2010.

Comparative evolutionary studies using ASE have already been performed in many organisms, for example in plants (Zhang and Borevitz, 2009), yeast (Tirosh et al., 2009) and animals (Wilson et al., 2008), including *Drosophila*. Actually, a rather large number of ASE studies between *D. melanogaster* and some of its closely related species already exist. In most of these studies expression was compared between *D. melanogaster* and *D. simulans* (Fontanillas et al., 2010; Graze et al., 2009, 2012; Landry et al., 2005; Wittkopp et al., 2004,

2008), one study compared *D. melanogaster* with *D. sechelia* (McManus et al., 2010) and the most recent one compared expression across the three species (Coolon et al., 2014). The methods used in these studies were very different (from pyrosequencing of a few selected genes to microarray and RNA-seq analysis), including the statistical analyses to infer differential gene expression, and therefore the results obtained are very different. However, all studies between *D. melanogaster* and *D. simulans* reported higher percentage of genes with divergent expression due to changes in *cis*, although in some cases only this type of regulation was studied (see Coolon and Wittkopp, 2013 for a review).

2.4.3 Regulatory divergence in developing tissues of three closely related *Drosophila* species

In all previously published ASE studies in *Drosophila*, adult tissue was analyzed (either whole animals or only heads) (Fontanillas et al., 2010; Graze et al., 2009, 2012; Landry et al., 2005; McManus et al., 2010; Wittkopp et al., 2004, 2008). Although gene expression divergence can influence morphological variation at all stages of an organism's life cycle, it is clear that the most important contribution takes place during development. It is during early stages of patterning of the body plan and the different tissues and organs that gene expression regulation is most important, and especially when the "toolkit" genes are active. In this study, I have used the three closely related species *D. melanogaster*, *D. simulans* and *D. mauritiana* (see also below 2.5.1) to try to better understand the mechanisms generating gene expression divergence at the early steps of species evolution. At a genome-wide level, these species need to have relatively conserved gene expression and GRN topology, since their head and eyes are extremely similar in morphology. However, some nodes of this network are divergent because they present, at least, significant differences in the size of their eyes and face (Arif et al., 2013; Hilbrant et al., 2014; Posnien et al., 2012; see also next section "**Eye size variation in two closely related *Drosophila* species**"). Thus, an ASE study can help to identify what type of regulatory mechanism is more widely present to generate expression differences between these species during head and eye development. To this aim I have performed the following inter-species crosses: *D. melanogaster* x *D. mauritiana* and *D. simulans* x *D. mauritiana*. In each case, I have dissected and sequenced the transcriptomes of eye-antennal imaginal discs of mid LIII and late LIII stage larvae from the F₁ hybrids of each of the two crosses and also from the three parental species. In order to study whether the results obtained are specific for this tissue for which significant differences in the

proportion of its organs have been described, I also sequenced the transcriptome of wing imaginal discs at mid LIII stage for both crosses and for all parents.

With this data, I aim to study the extent of gene expression divergence between three closely related species, two more closely related (*D. simulans* and *D. mauritiana*) and one slightly more distantly related (*D. melanogaster* diverged around 2.5 million years ago from the other two species (Lachaise et al., 1988)). First, I want to investigate if the major developmental processes that govern head and eye formation that I identify in my first project (“**New regulators governing *Drosophila* head development**”) are conserved even though these species show significant morphological differences in the size of their eyes and face (Arif et al., 2013; Hilbrant et al., 2014; Posnien et al., 2012). If that is the case, I will examine if there are genes in the underlying networks that have divergent gene expression in these three closely related species. Ultimately, I aim to classify each gene with divergent gene expression according to whether changes in *cis* or *trans* are responsible for the difference in orthologous gene expression. This comprehensive transcriptomics analysis in different developing tissues of closely related *Drosophila* species can provide new insights into the evolutionary mechanisms that govern gene expression divergence.

2.5 Eye size variation in two closely related *Drosophila* species

In order to reveal the molecular and developmental basis underlying morphological variation, the study of closely related species can be very useful. The short divergence time between species can increase the resolution of evolutionary studies (True and Haag, 2001), making it possible to identify the underlying locus and ideally even the causative mutation(s) that have generated the different phenotypes that can be observed in nature. There are already multiple publications of such micro-evo-devo studies where the genetic basis of morphological variation has been identified. For example, variations in the regulation of the gene *ultrabithorax* (*ubx*) have been found to give rise to different trichome patterns in the legs of different *Drosophila* species (Stern, 1998) and, more recently, differences in the regulation of the *unpaired-like* (*upd-like*) gene have been shown to modulate differences in wing size and shape between *Nasonia* wasp species (Loehlin and Werren, 2012). Here my aim is to identify the genetic basis of the variation in eye size observed between two very closely related species, *D. simulans* and *D. mauritiana*.

2.5.1 Eye size variation between *D. mauritiana* and *D. simulans*

The two studied species belong to the *melanogaster* subgroup (subgenus *Sophophora*, genus *Drosophila* (Bock and Wheeler, 1972; Sturtevant, 1939)), from which members are distributed mostly in Africa and the Asian-Pacific region. Besides *D. melanogaster*, *D. simulans* is one of the most extensively studied species and it was first described in the early 20th century (Sturtevant, 1919). This species, as *D. melanogaster*, is a cosmopolitan species that can be found all over the world, with the only exception of it being rare in East Asia. Most studies in *D. simulans* concentrated on comparing it to *D. melanogaster* in terms of e.g. population genetics, morphology, ecology or genome organization. *D. mauritiana* has only more recently been described (Tsacas and David, 1974) and it is endemic only in the Mauritius island, located east from Madagascar in the Indian Ocean, where neither *D. melanogaster* nor *D. simulans* are usually found. *D. simulans* and *D. mauritiana* are very closely related, and they diverged less than 0.5 million years ago (McDermott and Kliman, 2008). These two species are thought to have diverged from *D. melanogaster* approximately 2.5 million years ago (Lachaise et al., 1988).

Although these two species and *D. melanogaster* can only be reliably distinguished by their male genitalia morphology (Ashburner, 1989), other quantitative differences have been

described. For example, *D. melanogaster* has a broader cheek (face surface below the eye) than *D. simulans* (Burla, 1951) and *D. mauritiana* has the most dimorphic wings of the complete *melanogaster* subgroup (Gidaszewski et al., 2009).

A recent set of studies has shown that significant differences exist also in the size of the eyes and face of the closely related species *D. mauritiana* and *D. simulans* (Arif et al., 2013; Hilbrant et al., 2014; Posnien et al., 2012). *D. mauritiana* has significantly larger eyes compared to *D. simulans* (Figure 2.4.A). Conversely, *D. simulans* has a wider face than its sibling species. However, analysis of these differences during development of the eye-antennal imaginal discs indicate that these two traits are determined independently from each other (Arif et al., 2013). A more detailed comparison between the *D. mauritiana* TAM16 strain and *D. simulans* YVF strain showed that the observed differences in eye size are not due to a difference in the total number of ommatidia (Figure 2.4.B), but instead it is due to different ommatidia facet size (Figure 2.4.C).

2.5.2 Differences in ommatidia structure

In Posnien et al. 2012, the authors analyzed ommatidia size differences by measuring the area of the ommatidia facet, which is the hexagonal surface of the ommatidia (Ready et al., 1976). This facet corresponds to the corneal lens, and it is located at the top of the corresponding ommatidium. This measurement is a good indicator of eye size, since larger facets (if the total number of ommatidia is conserved) will inevitably result in larger compound eyes. However, the corneal lens provides little information about the structure of the underlying ommatidial cells, like the pigment cells and especially the photoreceptors, and about their organization (Hardy, 1985; Waddington, 1961). Importantly, these facet size differences could be caused by differences in the length of the ommatidia, the angle in which they are oriented or their width. A better insight into the underlying nature for the observed differences in the size of the ommatidial lens from *D. simulans* and *D. mauritiana* can also help in the understanding of the genetic basis of this variation.

Previous studies have imaged the inner structure of the *Drosophila* compound eyes using light microscopy, for example to study photoreceptor cluster organization (Reinke and Zipursky, 1988; Zheng et al., 1995) or to measure ommatidia length (Marrone et al., 2011) in mutant flies. This typically requires embedding of the adult heads in paraffin, resin or gel, followed by the very labor intensive process of micro-sectioning (Jenny, 2011; Tomlinson and Ready, 1987) and usually the analysis requires the use of an electron microscope. With these methods it is usually complicated to preserve the integrity of the analyzed sections. It

is especially challenging to compare size and shape measurements, since it is difficult to identify comparable sections between different individuals' heads. Furthermore, a 3D reconstruction from the obtained sections is not feasible, as each section is cut independently from the others.

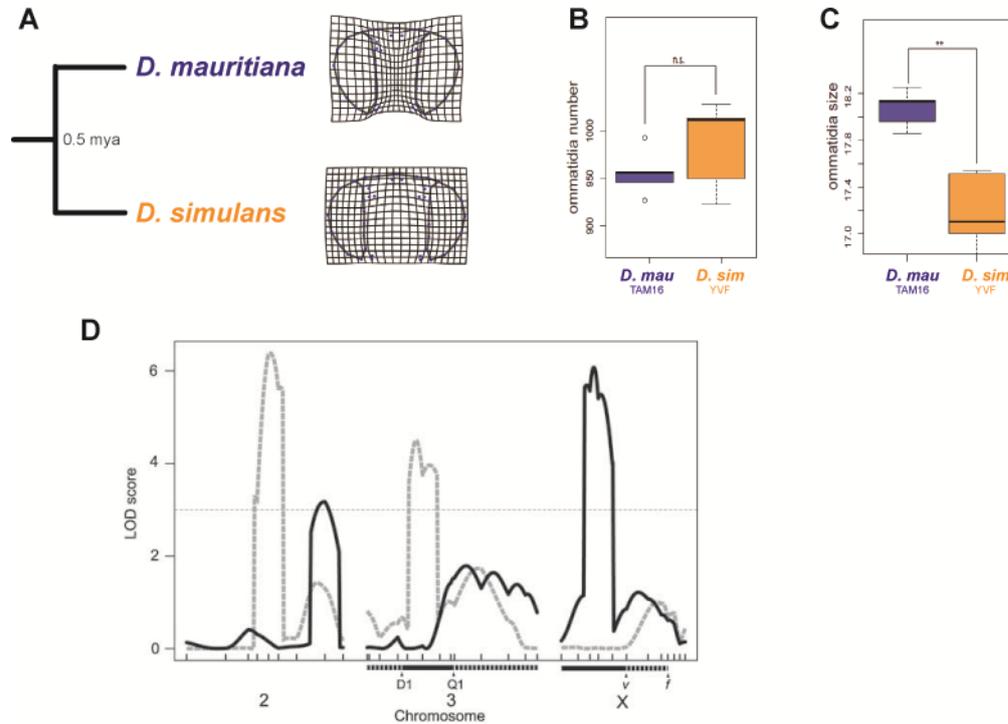


Figure 2.4. Eye and face size variation between *D. mauritiana* and *D. simulans*. (A) Adapted from Posnien et al. 2012. (B) The difference in number of ommatidia between *D. mauritiana* TAM16 and *D. simulans* YVF is not significantly different (Posnien et al. 2012). (C) The difference in the size of the ommatidia as measured by facet area is significantly larger in *D. mauritiana* TAM16 compared to *D. simulans* YVF (Posnien et al. 2012). (D) Quantitative trait loci related to eye size variation (black line) and face size variation (grey line) between the species *D. mauritiana* TAM16 and *D. simulans* YVF. Figure taken from Arif et al. 2013.

Here I have used a straightforward method to obtain high resolution images of the interior of intact *Drosophila* heads using confocal laser-scanning microscopy (McGurk et al., 2007; Smolla et al., 2014). This method consists of a complete clearing of all cuticular structures to make the interior accessible to the laser beam. Afterwards, cuticular structures inside the head can be imaged by virtue of their auto-fluorescence (Haug et al., 2011; Klaus et al., 2003). My purpose was to image the interior of the head of both *D. simulans* YVF and *D. mauritiana* TAM16 adult flies to take precise measurements of the ommatidia structure and compare these between the two species. First, I wanted to confirm the observed differences between ommatidia facet size and additionally I aimed to investigate if other features of the ommatidia of these species also show significant size differences.

2.5.3 A quantitative trait locus (QTL) correlates with eye size variation

Quantitative traits are those that show a continuous variation across populations and are likely to be controlled by multiple genes (Falconer and Mackay, 1995). Organ size is a clear quantitative trait since spatial measurements of any kind typically range in a continuous scale of values. Between closely related species a common method to identify causative mutations is quantitative trait loci (QTL) mapping. QTL analyses aim to identify the genetic basis of the variation in a trait by determining the genomic regions that are significantly linked together with a specific phenotype (Liu, 1998; Lynch and Walsh, 1998). This methodology depends on the availability of polymorphic molecular markers and the recent advances in molecular genetics techniques (Borevitz and Chory, 2004) has allowed the spread of high precision QTL analyses, mainly in the field of agriculture and farming (e.g. Baack et al., 2008; Frary et al., 2000; Hayes and Goddard, 2001), but also in medical, ecological and evolutionary studies (Cheverud and Routman, 1993; Erickson et al., 2004). However, this method usually identifies many different loci or very large genomic regions, so that the biggest challenge is to afterwards be able to identify the underlying causative polymorphism(s).

Here I have focused on the identification of the genetic causes of the variation in eye size between *D. mauritiana* TAM16 and *D. simulans* YVF. Recent data has identified different quantitative trait loci in the genome of *D. simulans* that are strongly associated with variation in compound eye size (Figure 2.4.D) (Arif et al., 2013). Using a combination of visible and molecular markers, different rounds of *D. mauritiana* x *D. simulans* crosses and F₁ back crosses were performed, followed by genotyping and phenotyping procedures to identify the region(s) in the genome that is/are significantly correlated with eye size differences. This analysis identified a region on the X chromosome as being highly associated with larger eyes (Figure 2.4.D, peak in black line). A second region with significant but with lower association score was identified in the 2nd chromosome. Interestingly, the loci significantly associated with face size (Figure 2.4.D, grey line) are different to the ones associated with eye size, supporting the developmental data that eye and face develop independently from each other (Arif et al., 2013).

Chromosomal introgression was also performed by Arif et al. 2013. This experiment confirmed that when *D. simulans* YVF had the identified X chromosome QTL region corresponding to *D. mauritiana* TAM16 genome, its eyes were bigger. The identification and usage of additional molecular markers allowed to increase the mapping resolution of this region that shows highest correlation with eye size to a region spanning only 1.1 Kb (7.4 -

8.5 Kb on the *D. simulans* X chromosome (genome assembly from Hu et al., 2013; Dr. Maria Santos Nunes, Oxford Brookes University, UK, unpublished results). Unfortunately, this region presents a high gene density and contains a total of 81 coding genes. At the start of my work, all these genes were putative candidate genes to be causing the observed differences in eye size between *D. mauritiana* TAM16 and *D. simulans* YVF. In order to reduce this list of candidate genes, I and Dr. Isabel Almudi (Oxford Brookes University, UK) performed several unbiased analyses to determine the genes that are more likely to be responsible for the observed eye size differences.

3 Materials and Methods

3.1 Fly strains, culture and crosses

Flies were kept on standard food at 25°C and 12h:12h dark:light cycle if not stated otherwise.

The *Drosophila* species used were *D. melanogaster* (OregonR), *D. mauritiana* (TAM16, collected in Mauritius in 2007 (Nolte et al., 2013)) and *D. simulans* (*yellow vermillion forked*, YVF; DSSC, University of California, San Diego, Stock no.14021-0251.146).

3.1.1.1 UAS/Gal4 crosses

For the Hunchback study I used the following fly lines: UAS-*hb*_{dsRNA} (Bloomington Stock Center #34704), *hb*-Gal4 (Vienna Tile library (Pfeiffer et al., 2008) VT038543, VT038544 and VT038545), UAS-*hb* (Bloomington Stock Center #8503), *repo*-Gal4/TM6B (kindly provided by Dr. Marion Sillies), UAS-stinger-GFP (nGFP) ((Barolo et al., 2000) kindly provided by Dr. Gerd Vorbrüggen), UAS-mCD8:GFP,UAS-H2B:RFP (kindly provided by the Wodarz Lab). Prof. Dr. Christian Klämbt kindly provided *moody*-Gal4 (Schwabe et al., 2005), Mz97-Gal4 (Ito et al., 1995) and c527-Gal4 (Ito et al., 1995). From Bloomington Stock Center I obtained all the lines expressing Gal4 under control of regulatory regions of the Hb putative target genes: *brk*-Gal4 (#53707), *CadN*-Gal4 (#49660), *Dl*-Gal4 (#45495), *Fas2*-Gal4 (#48471), *kni*-Gal4 (#50246), *rho*-Gal4 (#49379), *robo3*-Gal4 (#41256), *Sox21b*-Gal4 (#39803) and *Srv64B*-Gal4 (#49780). Additionally, I also used *Mef2*-Gal4 (#25756).

For the study on eye size variation I used the following lines: UAS-*CG1885*_{dsRNA} (Bloomington Stock Center, #51786), UAS-*Sptr*_{dsRNA} (VDRC (Dietzl et al., 2007), 17018/GD), GMR-Gal4 (Freeman, 1996) and UAS-*dicer* (Bloomington Stock Center, #36510).

All crosses were performed with an approximate ratio of 4:3 female:male flies. Crosses were always provided with additional yeast and were kept at 12h:12h dark:light cycle and controlled humidity, except the RNAi experiments, that were kept at 28°C and constant darkness.

3.1.1.2 *hb^{ts}* cross

hb^{ts1}, *rsd¹/TM3*, *Sb¹* flies (Bloomington Stock Center #1753) were crossed to *hb^{ts2}*, *st¹*, *e¹/TM3*, *Sb¹* flies (Bloomington Stock Center #1755) to generate a *hb^{ts1}/hb^{ts2}* stock. This line was kept at 18°C and constant light and larvae were only transferred to the restrictive temperature (28°C) for the loss of function experiments.

3.1.1.3 Inter-species crosses

400 *D. melanogaster* OreR or *D. simulans* YVF virgin females were crossed to 300 *D. mauritiana* TAM16 males respectively.

3.1.1.4 Dissections

Dissection time points are expressed as hours after egg laying (AEL) when eggs and larvae develop at 25°C. 72h AEL corresponds to late LII stage and 120h corresponds to late LIII stage (wandering larvae). These stages corresponded well in the three studied species. Mid LIII stage was defined as the time point when the morphogenetic furrow is located in the middle of the retinal field (analysis performed by Dr. Isabel Almudi, Oxford Brookes University, Oxford, UK). Only in *D. simulans* the advance of the differentiation wave was found to be slightly slower than the other two species. Therefore we dissected at 96h AEL in *D. melanogaster* and *D. mauritiana* but at 98h AEL in *D. simulans*. However, throughout this manuscript, this time point is referred to as “96h AEL”.

3.2 Immunohistochemistry

Antibody stainings were performed using standard procedures (Klein, 2008). Larvae were dissected in cold PBS solution and eye-antennal imaginal discs were dissected (together with the mouth parts and brain to facilitate the washing steps and to better preserve their integrity). All following steps were done in a rocking plate at room temperature. The dissected tissue was incubated in 4% paraformaldehyde solution for 25 minutes, followed by three washes with PBS + 0.3% Triton-X (PBT) and incubation in blocking solution (5% goat serum + 5% sheep serum in PBT) for 30 minutes. After that, the tissue was incubated with the corresponding primary antibodies for 90 minutes, followed by three more washes with PBT and incubation in blocking solution for 30 minutes. Next, the tissue was incubated with the corresponding secondary antibodies and/or Alexa Fluor 488 Phalloidin (Life Technologies, used at 1:100), then washed two times with PBT and then incubated

with DAPI (Carl Roth) for 10 minutes. The tissue was washed once more with PBT and after that with PBS, and finally transferred to mounting medium (80% glycerol + 4% n-propyl gallate in PBS) and left over-night at 4°C. The next day the tissue with mounting medium was transferred to a microscope slide and the discs were separated from the mouth parts and from the brain if necessary. Pictures were taken on a Zeiss LSM-510 confocal laser scanning microscope.

Antibodies used: rabbit α -Repo ((von Hilchen et al., 2013), 1:1000), guinea-pig α -Hb ((Kosman et al., 1998), 1:50), rabbit α -Hb (kind gift from Prof. Chris Q. Doe, 1:100), Cy3- α -HRP (kind gift from M Göpfert, 1:300), goat α -rabbit Alexa Fluor 488 (Invitrogen, 1:1000), goat α -rabbit Alexa Fluor 555 (Invitrogen, 1:100) and goat α -guinea-pig Alexa Fluor 555 (Invitrogen, 1:1000).

3.3 Blood-eye barrier assay

The integrity of the blood-eye barrier of *bunchback* knock-down flies was studied following the protocol from (Pinsonneault et al., 2011). *moody*-Gal4 virgin females were crossed with UAS-*bb*_{dsRNA} males at 28°C. UAS-*bb*_{dsRNA} flies were used as control and also raised at 28°C. 2-3 day old adults from these crosses were injected in the abdomen (Figure 4.1.14B) with 3-5 kDa FITC dextran (Sigma-Aldrich) (0.3 μ l the females and 0.2 μ l the males of 25 mg/ml solution). Animals were allowed to recover in fresh food over-night. Only surviving animals were scored. Dye penetrance in each eye was assessed qualitatively using a LEICA M205 FA fluorescent stereo microscope.

3.4 In situ hybridization

3.4.1.1 Molecular cloning

I cloned and performed *in situ* hybridization of 5 of the 14 reported candidate genes (see Results), namely *CG10958*, *CG1632*, *Sptr*, *sni* and *CG1885*. The remaining candidate genes were analyzed by Dr. Isabel Almudi, at the time member of the Research Group of Prof. Alistair McGregor in Oxford Brookes University (Oxford, UK).

D. simulans and *D. mauritiana* genomes were annotated as described in (Torres-Oliva et al. in revision). Annotated gene sequences of each pair of orthologs (Appendix 7.4) was aligned using MAFFT (Katoh et al., 2002). Primer3Plus software (Untergasser et al., 2007) was

used to design primers in a region with the minimal number of mismatches between the two species (Table 3.1). RNA was extracted from *D. simulans* LIII wandering larvae using RNeasy Mini Kit (Qiagen) and cDNA was synthesized using RevertAid First Strand cDNA Synthesis Kit (ThermoScientific). All genes were cloned into pCRII vector (Invitrogen) using standard techniques. Clones were sent for sequencing to LGC Genomics and the sequences were confirmed by local `blastn` to *D. simulans* reciprocally re-annotated transcriptome (Torres-Oliva et al. in revision).

Table 3.1. Primer sequences used to clone candidates from *D. simulans*.

Gene ID	Gene Name		Primers	Temp. [°C]	Sequence identity*
FBgn0030004	CG10958	<i>forward</i>	ATGCGGTGGAGAAGTGGCGC	61	99,4%
		<i>reverse</i>	TGAGGAAACCGCCGCGATCG		
FBgn0030027	CG1632	<i>forward</i>	GGACATCCGCTTCACCCGCC	61	98,2%
		<i>reverse</i>	CAGTGC GGCTTCTGGTGGCA		
FBgn0014032	Sptr	<i>forward</i>	TTGGCCGTGAGTTCGCCCAG	61	99,3%
		<i>reverse</i>	GCCGGGCGCGTAGTTCAACA		
FBgn0030026	sni	<i>forward</i>	TCGCGAGCAGGCAAAGGAGC	61	99,3%
		<i>reverse</i>	CGCCGTCTGTCTCTCGCCCA		
FBgn0030066	CG1885	<i>forward</i>	TATTCACATCGCCGCGCTGC	61	96,0%
		<i>reverse</i>	CAGATGCTCCACCGTCGGCC		

* similarity of the cloned sequence between *D. simulans* and *D. mauritiana*.

3.4.1.2 Probe synthesis

PCR was performed to amplify the correct fragments from the pCRII vector using M13 (GTAAAACGACGGCCAGTG) and M13 reverse (GGAAACAGCTATGACCAT) primers. The antisense *in situ* probes were then synthesized by *in vitro* transcription using Dig labeling mix (Roche) and T7 or Sp6 Polymerase (Roche) following the manufacturer's protocol.

3.4.1.3 Staining

D. simulans and *D. mauritiana* larvae were dissected at 120h AEL in PBS on ice. Discs attached to mouth parts were collected in an Eppendorf tube with ice-cold PBS for no longer than 20 minutes, when the PBS was replaced with 4% paraformaldehyde and incubated for 20 minutes on a rocking plate at room temperature. The tissue was washed three times with PBT (PBS + 0.1% Tween20) for 20 minutes. The following steps were performed at 65°C. The tissue was incubated in a 1:1 solution of PBT:Solution B (50% formamide + 5x saline-sodium citrate buffer + 0.1% Tween20) for 10 minutes, followed by two times 10 minutes in Solution B. Tissue was then pre-incubated for 10 minutes in

Solution A (100 µg/ml denatured herring sperm DNA + 50 µg/ml heparin in Solution B) followed by a 1h incubation in Solution A. The DIG-labelled probe was then diluted in Solution A (0.5 µl of probe in 100 µl Solution A) and hybridized over-night. The tissue was then rinsed in a graded series of Solution B:PBT (3:1, 1:1, 1:3). The tubes were transferred back to room temperature for the following steps. The tissue was washed two times for 20 minutes in PBT and subsequently incubated 20 minutes in blocking solution (0.14 g albumin fraction V + 280 µl sheep serum + 280 µl goat serum in PBT) and then incubated for 90 minutes with the anti-DIG antibody (1:2000 in blocking solution, Sigma-Aldrich). After that, the tissue was washed three times in PBT for 20 minutes and then rinsed with AP buffer (100 mM NaCl + 500 mM MgCl₂ + 100 mM Tris-HCl at pH 9.5 + 0.1% Tween20) three times for 5 minutes. Finally, the reaction mix (4,5 µl NBT + 3,5 µ BCIP + 1 ml AP buffer-tween) was added and the discs were transferred to a block dish to control the staining. The discs were stained for approximately 3h depending on the probe. Note that stainings for the same gene were stopped at the same time in the two species. The reaction was stopped by washing three times with PBT and then the mounting medium (80% glycerol) was added. Discs were then transferred to a microscope slide and prepared by separating them from the mouth parts. Pictures were taken with a Zeiss Axioplan 2 microscope.

3.5 Optical sectioning of *Drosophila* heads

Fly heads were cleared following the protocol from (Smolla et al., 2014). *D. simulans* and *D. mauritiana* flies were raised separately at 25°C in 12h:12h light:dark cycle. 5 days old flies were anesthetized with CO₂, their heads cut and placed in 4% paraformaldehyde and left over-night at 4°C. Second left legs of each individual were also dissected and kept on sticky tape to estimate whole body size. Heads were washed three times with PBT and then transferred to 15% H₂O₂ solution to remove eye pigmentation. After 5 days the depigmented heads were washed 3 times with PBS and then dehydrated in a graded ethanol series (50%, 70%, 90%, 95% and three times 100%). Finally, heads were cleared in methylsalicylate. Cleared heads were mounted on microscope slides facing up and covered with cover slides using modelling clay spacers and applying no pressure to prevent flattening. Pictures were taken on a Zeiss LSM-510 confocal laser scanning microscope with 488 nm emission light (argon laser) using a 20x (0.5 NA) dry objective and 0.8 zoom. Heads were scanned from top to bottom in 50 sections, of approximately 5 µm each. Head

reconstruction from the 50 sections was made using AMIRA 3D software v4.5.4 (FEI Company, Berlin, Germany).

Measurements were taken on the central section of the heads for both species. To make sure that the section was the same in all individuals, the first section where the lamina was clearly visible beginning from the dorsal side was selected and used to take all measurements (Figure 4.4.3B). Measurements were made with Fiji (Schindelin et al., 2012). For ommatidia length (from the base of the ommatidia and from the base of the lens) and ommatidia width (from the pseudocones and from the lens), the measurement was taken for the 5 central ommatidia of each analyzed eye and the mean of these 5 measurements was used. To correct for body size, the residuals of the linear regression between the measured trait and the length of the tibia of the second left leg were eventually compared. T-test was applied to determine if the differences between species were significant.

3.6 RNA-seq and bioinformatics analysis

3.6.1.1 RNA extraction and sequencing

RNA-seq of *Drosophila* larval imaginal discs was performed for all the described projects, but using different stages (72h AEL, 96h AEL and/or 120h AEL), tissues (mostly eye-antennal imaginal discs but also wing imaginal discs) and species (*D. melanogaster*, *D. simulans* and/or *D. mauritiana*) and different analysis pipelines according to the biological question I wanted to answer (Table 3.2). The procedures for sample preparation and sequencing were the same in all cases (if not stated otherwise) and are described here first. The project-specific details and analysis steps are described below for each section. The description of the methods for the reciprocal (re)-annotation of closely related species' genomes is included in the respective manuscript (section 4.2).

Table 3.2. Summary of RNA-seq samples.

Species	Stage	Tissue	Sex	Type	Project*
<i>D. melanogaster</i>	72h	eye	male+female	SE 50 bp	A, C
	96h	eye	female	SE 50 bp	A, C
	120h	eye	female	SE 50 bp	A, B, C
	96h	wing	male+female	SE 50 bp	C
<i>D. mauritiana</i>	72h	eye	male+female	PE 100 bp	C
	96h	eye	female	SE 50 bp	C
	120h	eye	female	PE 100 bp	B, C, D
	120h	eye	female	SE 50 bp	B, C

	96h	wing	male+female	SE 50 bp	C
<i>D. simulans</i>	72h	eye	male+female	PE 100 bp	C
	96h	eye	female	SE 50 bp	C
	120h	eye	female	PE 100 bp	B, C, D
	96h	wing	male+female	SE 50 bp	C
<i>D. melanogaster</i> × <i>D. mauritiana</i>	96h	eye	female	SE 50 bp	C
	120h	eye	female	SE 50 bp	C
	96h	wing	male+female	SE 50 bp	C
<i>D. simulans</i> × <i>D. mauritiana</i>	96h	eye	female	SE 50 bp	C
	120h	eye	female	SE 50 bp	C
	96h	wing	male+female	SE 50 bp	C

***A**: New regulatory interactions governing *Drosophila* head development; **B**: Torres-Oliva et al. in revision; **C**: Gene expression divergence in closely related *Drosophila* species; **D**: Eye size variation between two closely related *Drosophila* species. “SE”: single-end reads; “PE”: paired-end reads.

Parental flies were raised at 25°C and 12h:12h dark:light cycle for at least two generations and their eggs were collected in 1h windows. Freshly hatched LI larvae were transferred into fresh vials in density-controlled conditions (30 freshly hatched LI larvae per vial). At the required time point, either only female larvae (for the 96h and 120h eye-antennal imaginal disc samples) or male and female larvae (for the 72h eye-antennal imaginal disc samples and the 96h wing disc samples) were dissected and eye-antennal/wing discs were stored in RNALater (Qiagen, Venlo, Netherlands). We dissected 40-50 discs for the 120h samples, 80-90 discs for the 96h samples and 120-130 discs for the 72h samples. We generated three biological replicates for each sample type. This procedure was performed by Dr. Isabel Almudi (Oxford Brookes University, UK), the Master students Elisa Buchberger and Melissa Jüds and me.

The following steps were performed by the Transcriptome and Genome Analysis Laboratory (TAL) in Göttingen. Total RNA was isolated using the Trizol (Invitrogen, Thermo Fisher Scientific, Waltham, Massachusetts, USA) method according to the manufacturer’s recommendations and the samples were DNaseI (Sigma, St. Louis, Missouri, USA) treated in order to remove DNA contamination. RNA quality was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) microfluidic electrophoresis. Only samples with comparable RNA integrity numbers were selected for sequencing.

Library preparation for RNA-seq was performed using the TruSeq RNA Sample Preparation Kit (Illumina, catalog ID RS-122-2002) starting from 500 ng of total RNA. Accurate quantitation of cDNA libraries was performed using the QuantiFluor™dsDNA System (Promega, Madison, Wisconsin, USA). The size range of final cDNA libraries was

determined applying the DNA 1000 chip on the Bioanalyzer 2100 from Agilent (280 bp). cDNA libraries were amplified and sequenced using cBot and HiSeq 2000 (Illumina): only *D. simulans* and *D. mauritiana* 120h eye-antennal imaginal disc samples were sequenced as paired-end (PE) reads (2 x 100 bp), all the rest of samples were sequenced in single-end (SE) reads (1 x 50 bp). Sequence images were transformed to bcl files using the software BaseCaller (Illumina). The bcl files were demultiplexed to fastq files with CASAVA (version 1.8.2).

3.6.1.2 Quality control

I carried out quality control analysis using FastQC software (version 0.10.1, Babraham Bioinformatics). I identified a number of samples coming from the same lane with a peak of N bases in the same position, probably a product of the presence of bubbles in the sequencing plate. These samples were re-sequenced and this phenomenon disappeared. All samples had Phred quality score >Q10 and only few had <Q20. Following recently published guidelines (Macmanes, 2014; Williams et al., 2016) I did not trim these bases, but instead relied on the aligner software to make the quality call.

At present, only the samples used in Torres-Oliva et al. (in revision) (*D. melanogaster* 120h (SE 50 bp), *D. mauritiana* 120h (SE 50 bp and PE 100 bp) and *D. simulans* 120h (PE 100 bp)) have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE76252 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76252>).

3.6.2 New regulatory interactions governing *Drosophila* head development

The RNA-seq reads used for this analysis were *D. melanogaster* OregonR eye-antennal imaginal discs at 72h AEL, 96h AEL and 120h AEL (all SE 50 bp, 3 biological replicates for each stage). I downloaded the transcript sequences (only CDS) of *D. melanogaster* (r5.55) from FlyBase and used a python script (kindly provided by Nicola Palmieri (University of Veterinary Medicine, Vienna)) to extract only the longest transcript per gene. This sequences were used as reference to map the reads using Bowtie2 (Langmead and Salzberg, 2012) with parameters `-very-sensitive-local -N 1`. The number of reads mapping to each transcript were summarized using the command `idxstats` from SAMtools v0.1.19 (Li et al., 2009).

3.6.2.1 Differential expression analysis and GO term enrichment

For each pair-wise comparison (72h vs. 96h and 96h vs. 120h) HTSFilter (Rau et al., 2013) was used with default parameters to filter out genes with very low expression in all samples. For the remaining genes in each pair-wise comparison, differential expression was calculated using DESeq2 v1.2.7. with default parameters (`design = ~ time`) (Love et al., 2014a). Gene Ontology (GO) terms enrichment analysis for Biological Process was performed with GO TermFinder (Boyle et al., 2004) with default parameters. Only the first non-redundant terms were plotted.

3.6.2.2 Expression clusters

In this analysis, the count data for the three time points (72h, 96h and 120h AEL) was used. HTSFilter (Rau et al., 2013) was again used to discard lowly expressed genes across all samples. The function `PoisMixClusWrapper` from the library HTSCluster (Rau et al., 2015) was applied on the rest of genes with the parameters: `gmin=1`, `gmax=25`, `lib.type="DESeq"`. Genes with predicted MAP < 99% were discarded. For the plots, the variance stabilizing transformation from DESeq2 (Love et al., 2014a) library was used to normalize the background read count of the genes belonging to each cluster.

The GO terms enriched in each cluster of genes were obtained with the plugin BiNGO (Maere et al., 2005) in Cytoscape v3.1.1 (Cline et al., 2007) in batch mode and default parameters. Only the first four non-redundant terms are reported. The dataset with all known genetic interactions among *Drosophila* genes was downloaded from DroID (data version 2014_10) (Murali et al., 2011). Networks of genetic interactions between genes of each cluster were constructed with Cytoscape.

The transcription factors enriched to regulate the genes of each cluster were obtained with i-cisTarget method (Herrmann et al., 2012) with the following parameters: dm3 assembly, only "TF binding sites", 5 Kb upstream and full transcript as mapping region, 0.4 as minimum fraction of overlap, 3.0 as NES threshold and 0.01 ROC threshold.

3.6.2.3 Targets analysis

From Berkeley *Drosophila* Transcription Network Project (BDTNP) site (Li et al., 2008), I downloaded BED files for the Hb (anti-Hb (antibody 2), stage 9) CHIP-chip experiment (Symmetric-null test and 1% FDR cutoff). The LiftOver tool from UCSC Browser (Kent et al., 2002) was used to transform the dm2 coordinates into the dm3 assembly. The closest gene to each CHIP-chip interval was identified with the script `annotatePeaks.pl` from

the HOMER suite of tools (Heinz et al., 2010). I confirmed that the regulatory regions of the identified genes were enriched for the Hb motif with the script `findMotifGenome.pl` from the same suite. I checked then in which of the closest genes to the ChIP-chip intervals the Hb motif (searched as matrix) could be identified using again the script `annotatePeaks.pl` with the parameters `tss -size -1000,1000 -m motif_matrix`. The genes with at least one instance of the motif were selected as Hb high confident targets.

3.6.3 Gene expression divergence in closely related *Drosophila* species

All reads summarized in Table 3.2 were used for this analysis. These samples were obtained from different dissections at different days (October 2012, August 2013 and November 2015). The use of different sequencing types corresponds to these different experiments: in October 2012 we generated PE 100 bp reads and in the other experiments SE 50 bp. Before the mapping step, PE 100 bp reads were converted into SE 50 bp by splitting the reads in half and merging right and left reads into a single file.

To analyze the possible bias due to sequencing type, a matrix with the count values for the parental samples (eye-antennal imaginal discs at 72h, 96h and 120h AEL) of the three species was normalized using `normalizeQuantiles` from the `limma` package (Ritchie et al., 2015), then log transformed and `plotMDS` from the same package was used on the resulting matrix.

3.6.3.1 Differential expression between stages of three *Drosophila* species

The reciprocally re-annotated references described in Torres-Oliva et al. (in revision) were used to map the species-specific reads. Bowtie2 (Langmead and Salzberg, 2012) was used to map the reads to each reference (`-very-sensitive-local -N 1`) and the `idxstats` command from SAMtools v0.1.19 (Li et al., 2009) was used to summarize the number of mapped reads.

HTSFilter (Rau et al., 2013) was used to discard lowly expressed genes across all samples. On the rest of genes, HTSCluster (Rau et al., 2015) was applied to cluster genes by expression pattern with the function `PoisMixClusWrapper` with the parameters: `gmin=1, gmax=25, lib.type="DESeq"`. The genes in clusters 1 and 2 were separated and re-clustered using the same parameters (see Results). Clusters 3 to 8 and all subclusters (resulting from the clustering of genes in clusters 1 and 2) were introduced to BiNGO (Maere et al., 2005) to identify enriched GO Terms and to *i-cisTarget* (Herrmann et al.,

2012) to identify transcription factors enriched as regulators of the genes in each cluster (parameters: only “TF binding sites”, 5 Kb upstream and full transcript as mapping region, 0.4 as minimum fraction of overlap, 4.0 as NES threshold and 0.01 ROC threshold).

The count data for all the samples (eye-antennal imaginal discs at 72h, 96h and 120h AEL) was also analyzed for differential gene expression using DESeq2 v.1.2.7 (Love et al., 2014a), specifying `design = ~ species + time` in the `DESeqDataSetFromMatrix` function. The order of the levels of the `time` factor was specified to be `72h, 96h, 120h`, so that the calculated variation was between 72h and 120h. The 1,000 most differentially expressed genes (defined by lowest p-adjusted values) were selected and their normalized counts values obtained (see above: `normalizeQuantiles`). The distances between each gene (row) were calculated with the `dist` function (R Core Team, 2015) (`method = "euclidean"`) and then hierarchical clustering analysis was performed with the `hclust` function (`method = "complete"`). The resulting dendrogram was used to separate the data into 8 clusters with the function `cutree(k = 8)`. Finally, the results were plotted with the `heatmap.2` function. The genes belonging to each cluster were analyzed with the `i-cisTarget` method (Herrmann et al., 2012) as described in section 3.6.2.2.

3.6.3.2 Differential expression between species

For this analysis, the same raw counts were used as in the previous section. However, nine pair-wise analyses were performed, one between each pair of species for each time point using DESeq2 (Love et al., 2014a) with default parameters (`design = ~ species`). Afterwards the complete dataset (3 species, 3 time points) was analyzed with DESeq2 using `design = ~ time + species`. The order of the levels of the `species` factor was specified to be `D. mel, D. mau, D. sim`, so that the calculated variation was between *D. melanogaster* and *D. simulans*. The same procedure as described previously was followed to cluster the 1,000 most differentially expressed genes into 8 clusters. In `i-cisTarget`, the collection of PWMs was searched for enrichment as well as the collection of ChIP-chip experiments for transcription factor enrichment.

3.6.3.3 Generation of strain-specific references

To generate strain specific genomes, first genomic high throughput data was obtained for the strains of interest: *D. mauritiana* TAM16 and *D. simulans* YVF reads were kindly provided by Dr. Alistair McGregor (PE 72 bp, generated with Illumina GenomeAnalyzer IIx, unpublished) and *D. melanogaster* OreR reads were downloaded from the Sequence

Read Archive using the SRA toolkit v2.5.2 (experiment SRX671606, run SRR1538754, PE 125 bp, generated with Illumina GenomeAnalyzer IIX by the Baylor College of Medicine as part of the modENCODE project). Reads were trimmed using the script `trim-fastq.pl` from the PoPoolation software v1.2.2. (Kofler et al., 2011) with parameters `-quality-threshold 18 -min-length 50`. Trimmed reads were aligned to the corresponding published genome: OreR reads were mapped to the reference *D. melanogaster* M36 genome r5.55 (Tweedie et al., 2009), *D. mauritiana* TAM16 reads were mapped to the published *D. mauritiana* MS17 genome (Nolte et al., 2013) and *D. simulans* YVF reads were mapped to the reference *D. simulans* w⁵⁰¹ genome (Hu et al., 2013). Reads were mapped using the `aln` command from BWA v0.7.5 (Li and Durbin, 2009) with parameters `-l 150 -o 2 -d 12 -e 12 -n 0.01`, followed by the `sampe` command from the same program. SAMtools v0.1.19 (Li et al., 2009) was used to discard reads that did not map unambiguously or that were not correctly paired (`view -q 20 -F 0x0008 | view -F 0x0004 | view -f 0x0002`). To exclude possible contamination, the genomes of *Wolbachia* (Genebank accession number NC 006833.1), *Acetobacter pasteurianus* (AP011121.1) and *Lactobacillus plantarum* (AL935263.2) (all downloaded from NCBI) were also used as references and reads mapping to them were discarded. The `mpileup` command from SAMtools was used to list all the variants (single nucleotide polymorphisms (SNPs) and insertions and deletions (INDELs)) between the reference genomes and the specific strains used in this study. A python script kindly provided by Dr. Martin Kapun (University of Lausanne, Switzerland) was used to replace the high confidence variants present in the strain-specific genome in the reference genomes (`mpileup_parse.py -base-quality-threshold 20 -coverage-threshold 5,20`).

The strain-specific genomes were reciprocally re-annotated as described in Torres-Oliva et al. (in revision) but using the longest *D. melanogaster* full transcript sequences (instead of CDS sequences) in order to include UTR regions (r5.55, downloaded from the FlyBase site).

The references for the *D. simulans* x *D. mauritiana* F₁ hybrids analysis were not specific enough (see Results, section 4.3.3.1) and a second round of re-assembly was applied, in this case using the parental RNA-seq reads in the strain-specific, reciprocally re-annotated transcripts set (described in the previous paragraph). Reads from the first replicate of our *D. simulans* YVF 96h (eye-antennal and wing discs) and 120h eye-antennal imaginal disc samples were concatenated and mapped against the generated YVF transcript set; in parallel, reads from the first replicate of my *D. mauritiana* TAM16 96h (eye-antennal and

wing discs) and 120h eye-antennal imaginal disc samples were concatenated and mapped against the generated TAM16 transcript set. Reads were aligned using BWA with parameters `-l 150 -o 2 -d 12 -e 12 -n 0.01`, followed by the `samse` command. SAMtools (Li et al., 2009) was used to keep only the unambiguously mapped reads (`view -F 0x0004`). The detected SNPs and INDELS were summarized with the `mpileup` command from SAMtools. Since in RNA-seq experiments coverage depends on gene expression and it can be extremely high, the python script from Dr. Martin Kapun was modified not to require a maximum coverage value.

To check if the number of SNPs between species increased with the newly generated references, mismatches in the original published references and in the corresponding newly generated strain-specific references were detected with `blastn` between orthologous genes.

3.6.3.4 Detection of allele-specific expression and inference of regulation type

For this analysis, RNA-seq reads from the parental samples were mapped to the corresponding species reference. In contrast, in order to detect allele-specific expression, RNA-seq reads from the F₁ hybrid individuals were mapped to a combined reference of both parental species (Figure 3.1).

To map the reads to the corresponding references, Bowtie v1.0.0 (Langmead et al., 2009) was used for all samples. This version (instead of Bowtie2) was preferred due to the available `-best -strata` mode, which classifies the read alignments in different “strata”, according to the number of mismatches. This is of key importance in the analysis of reads coming from F₁ hybrid animals, since one can only allow the reads to be reported as mapped if there is only one best possible alignment, meaning that at least one polymorphism differentiates the reference sequence between the two parental species at that position. The reads from parental samples were mapped to the species-specific reference transcriptomes, i.e. reads from *D. simulans* YVF were mapped to the newly generated RNA-seq-based YVF reference, *D. melanogaster* OreR reads were mapped to the newly generated DNA-seq-based OreR reference and *D. mauritiana* were mapped twice independently: for the comparison with *D. melanogaster*, the reads were mapped to the first version of the TAM16 specific reference (based on DNA-seq), and for the comparison with *D. simulans*, the reads were mapped to the RNA-seq-based version of the TAM16 specific reference. For the comparison between *D. melanogaster* and *D. mauritiana*, the Bowtie parameters used were `-S -best-strata -v 1 -m 1`, where the option `-v 1` allows one mismatch between the mapped read and the reference sequence; since these two

species showed enough differentiating polymorphisms, this option can increase the number of mapped reads. However, for the comparison between *D. simulans* and *D. mauritiana*, the Bowtie parameters used were `-S -best-strata -v 0 -m 1`, where no mismatch is allowed in order to increase specificity. The reads from hybrid samples were mapped to a reference that contained the sequences of both parental species, to represent the two possible alleles. In this FASTA file, the header of each allele included a label to indicate the corresponding species. Reads from *D. melanogaster* x *D. mauritiana* hybrids were mapped with Bowtie with parameters `-S -best-strata -v 1 -m 1`, and reads from *D. simulans* x *D. mauritiana* hybrids were mapped with parameters `-S -best-strata -v 0 -m 1`. Read counts were summarized with the command `idxstats` from SAMtools (Li et al., 2009).

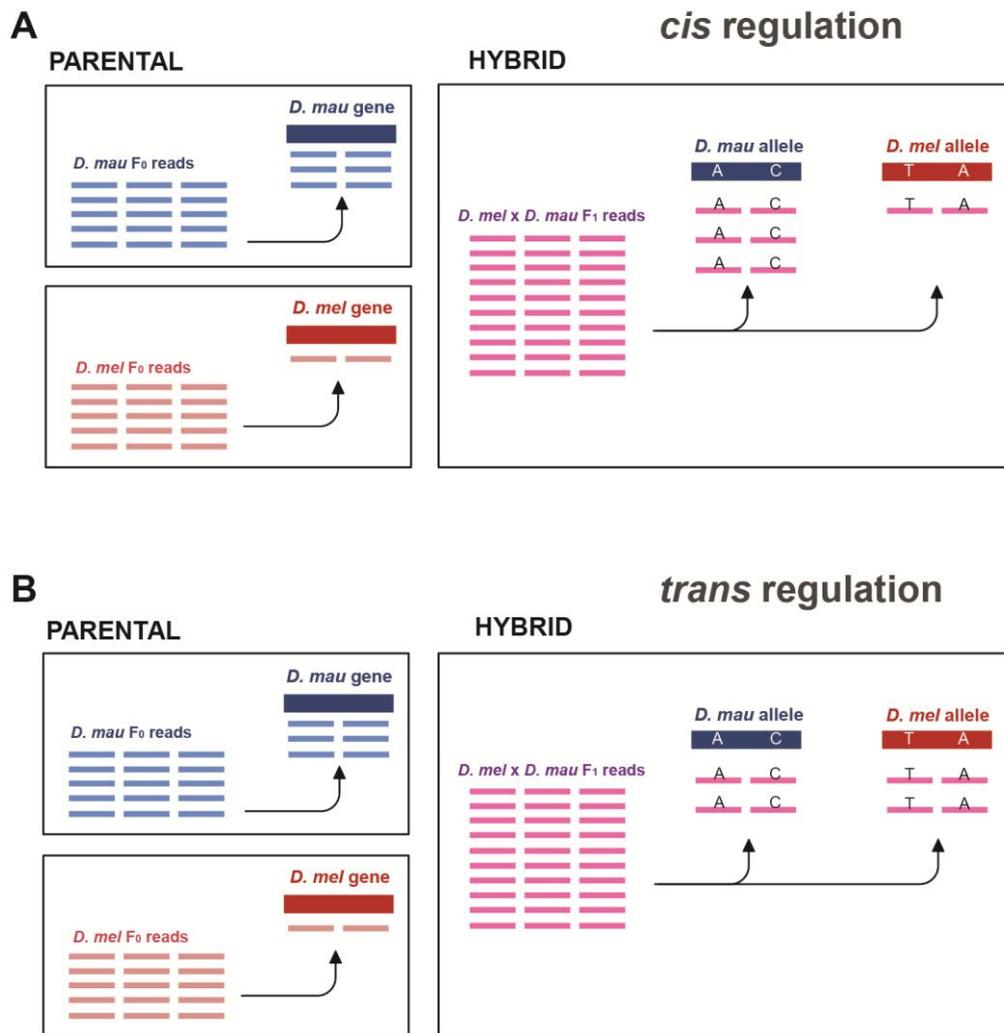


Figure 3.1. Mapping of parental and hybrid reads. Parental reads are mapped to the species-specific reference (panels on the left). Reads from the hybrid individuals are mapped to a reference that contains the reference of both parents (panels on the right). Polymorphisms that distinguish the reference from each species are used to map the reads to the correct allele. Only reads that map uniquely to one of the two alleles are kept. **(A)** and **(B)** show examples of the analysis of a gene with divergent expression levels, which is higher expressed in *D. mauritiana* individuals. **(A)** If the *D.*

mauritiana allele is also higher expressed than the *D. melanogaster* allele in hybrid individuals, it means that the expression divergence is due to variation in the *cis* regulatory region of the gene. **(B)** If the *D. mauritiana* and *D. melanogaster* alleles are equally expressed in the hybrid individuals, it means that the divergent expression detected in the parents is due to variation in upstream factors (*trans*).

The analysis of regulatory divergence was performed in R v3.1.2 (R Core Team, 2015). Differential expression between the parental samples (*D. melanogaster* vs. *D. mauritiana* and *D. simulans* vs. *D. mauritiana*) was analyzed with DESeq2 (Love et al., 2014a) and default parameters (`design = ~ species`). The count-tags from the hybrid samples were first split according to which of the two parental species' allele they corresponded to (using the label present in the headers). Then an allele-specific differential expression analysis was carried out with DESeq2 (Love et al., 2014a) with default parameters. The `overLapper` function from `systemPipeR` library (Girke, 2015) was used to identify the overlap of genes differentially expressed in the parental and/or in the hybrids. Genes not differentially expressed between parental species nor hybrids were considered to have conserved expression; genes differentially expressed in the parents and in the hybrids with the same species having higher expression in both cases were considered to have divergent expression because of variation in *cis* (Figure 3.1A); genes differentially expressed in the parents but not in the hybrids were considered to be divergent because of variation in *trans* (Figure 3.1B); genes differentially expressed in the parents and in the hybrids but in opposite direction were considered to have *cis* \times *trans* regulation; genes differentially expressed in the hybrids but with conserved expression in the parental species were considered to have compensatory regulation (McManus et al., 2010; Wittkopp et al., 2004).

3.6.4 Eye size variation in two closely related *Drosophila* species

The RNA-seq reads used for this analysis were *D. simulans* YVF and *D. mauritiana* TAM16 eye-antennal imaginal discs at 120h AEL (Table 3.2). All samples were PE 100 bp and, prior to mapping, were converted into SE 50 bp by splitting the reads in half and merging right and left reads into a single file. Reads were mapped to reciprocally re-annotated (Torres-Oliva et al. in revision) strain-specific references (generated in Oxford Brookes University, unpublished) using Bowtie2 (Langmead and Salzberg, 2012) (`-very-sensitive-local -N 1`) and the total number of reads mapping to each transcript were obtained with the command `idxstats` from SAMtools v0.1.19 (Li et al., 2009).

3.6.4.1 Differential expression analysis

Genes with less than 1 read per million reads in at least 4 samples were filtered out before starting the differential expression analysis. Differential expression analysis was performed with two programs with default parameters: edgeR (Robinson et al., 2010) and DESeq (Anders and Huber, 2010). Although an improved version of DESeq (namely DESeq2) is currently available, this analysis was performed in 2013, when the new software was not yet published. The candidates that were later tested were based on this initial analysis; therefore I report here the results for DESeq (first version) although in the other sections of the Thesis I have used DESeq2.

3.6.4.2 Analysis of coding sequence identity

The genomic references of each strain (generated in Oxford Brookes University, unpublished) were annotated using the longest transcript isoform (only CDS) from *D. melanogaster* r5.55. For this annotation, Exonerate v2.2 (Slater and Birney, 2005) was used with parameters—`model est2genome—softmasktarget yes—bestn 1 --minintron 20 --maxintron 20000`. The species-specific transcript sequence of the genes on the QTL region which are expressed in eye-antennal imaginal discs (76 genes) were obtained from these annotations. Geneious v6.0.6 (Kearse et al., 2012) was used to translate the transcript sequences and MAFFT v7.017 (Katoh et al., 2002) was used to align each pair of orthologs.

4 Results

4.1 New regulatory interactions governing *Drosophila* head development

4.1.1 Differentially expressed genes during head development

To identify the genes expressed during *D. melanogaster* eye-antennal imaginal disc development and their expression dynamics, I performed RNA-seq on this tissue at three relevant larval stages: 72h AEL (late LII; before the process of photoreceptor differentiation has started in the eye disc), 96h AEL (mid LIII; when the morphogenetic furrow is in the middle of the retinal field) and 120h AEL (late LIII; at the end of morphogenetic furrow progression).

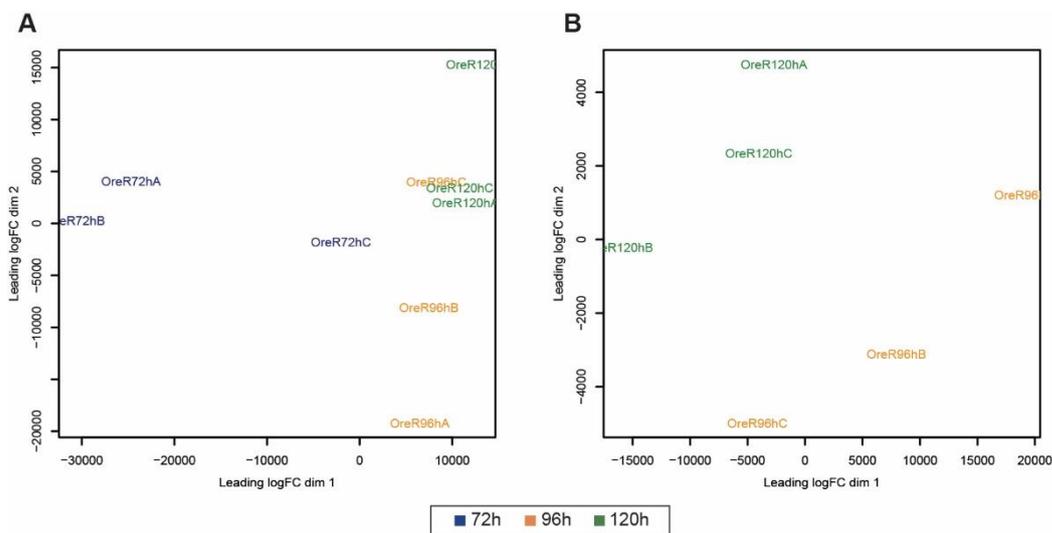


Figure 4.1.1. Multidimensional scaling plot of *D. melanogaster* samples. The 9 studied samples are represented with a label “OreR” + *time point*. Labels are color coded according to the corresponding time point. **(A)** First two dimensions separating the 9 samples from the three stages (three replicates per stage). **(B)** First two dimensions separating 96h and 120h samples. 72h samples are removed from the analysis.

I sequenced three biological replicates of each time point, and once the reads had been mapped to the *D. melanogaster* reference, I analyzed the count matrices by multidimensional scaling clustering (Fig. 4.1.1A). Dimension 1 clearly separates the 72h replicates from the later stages indicating that the largest difference in gene expression is between LII eye discs (72h) and LIII eye discs (96h and 120h). Dimension 2 separates samples 96h A and B from 120h B. However, samples 96h C, 120h A and 120h C cluster together. This scaling method uses the genes with largest fold change differences to separate the data, but it is

likely that the genes varying between 72h and the other two time points are not changing much between 96h and 120h. Therefore, I repeated the analysis without the 72h samples (Fig. 4.1.1B). In this case, the two dimensions can separate the data better, with the 96h samples situated on the lower right corner of the two-dimensional plot and the 120h on the top left corner.

Before performing differential expression analysis, I removed the genes that were not expressed or that had very low expression levels across all samples. These genes are unlikely to be differentially expressed and, additionally, their discreteness interferes with the statistical analysis of genes with larger expression levels. Filtering of lowly expressed genes with HTSFilter (Rau et al., 2013) indicated that 9,020 genes are expressed at 72h and/or 96h, and that 8,134 genes are expressed at 96h and/or 120h (Table 4.1.1). I used DESeq2 (Love et al., 2014a) to identify the genes that are differentially expressed during the studied developmental transitions. As anticipated by the multidimensional scaling plot (Figure 4.1.1A), the number of genes that change their expression between 72h and 96h is much larger than between 96h and 120h (Table 4.1.1). In only 24 hours, during the transition from LII to LIII, 50% of the expressed genes change their expression significantly. In the transition from 96h to 120h, in contrast, only 22% of the genes undergo a change in their expression.

Table 4.1.1. Differentially expressed genes.

Time points	Expressed genes	Up-regulated	Down-regulated
72h vs. 96h	9,020	2,897 (32.12%)	2,591 (28.72%)
96h vs. 120h	8,134	898 (11.04%)	887 (10.90%)

Expressed genes are those that passed the HTSFilter cut-off. Differentially expressed are those with $p\text{-adj} < 0.05$.

To better understand and characterize these developmental transitions, I investigated the biological processes that are involved in each stage. To that aim I performed a Gene Ontology (GO) term enrichment analysis on the genes that are differentially expressed in each transition. I found 1,010 GO terms enriched for the genes that are up-regulated between 72h and 96h discs and only 99 GO terms on the down-regulated genes. In the transition from 96h to 120h, 113 GO terms were enriched on the up-regulated genes and 71 on the down-regulated.

The transition from LII to LIII represents a complete shift in the biological processes that are taking place in the eye-antennal imaginal disc (Figure 4.1.2A). The genes that are down-regulated are mostly related to metabolism and energy production. The up-regulated genes,

in contrast, code for transcription factors (regulators of biological process and of gene expression) and other proteins involved in differentiation, signaling, growth, axon projection and eye development, among others. Interestingly, these are also some of the categories that are still enriched in the genes that have higher expression at 120h compared to 96h (Figure 4.1.2B). In this transition, the genes up-regulated are also involved in signaling, regulation and differentiation, and also in chemotaxis, pigmentation and R7 cell differentiation. The down-regulated genes are mainly involved in cell cycle processes, as well as generation of energy and cuticle development.

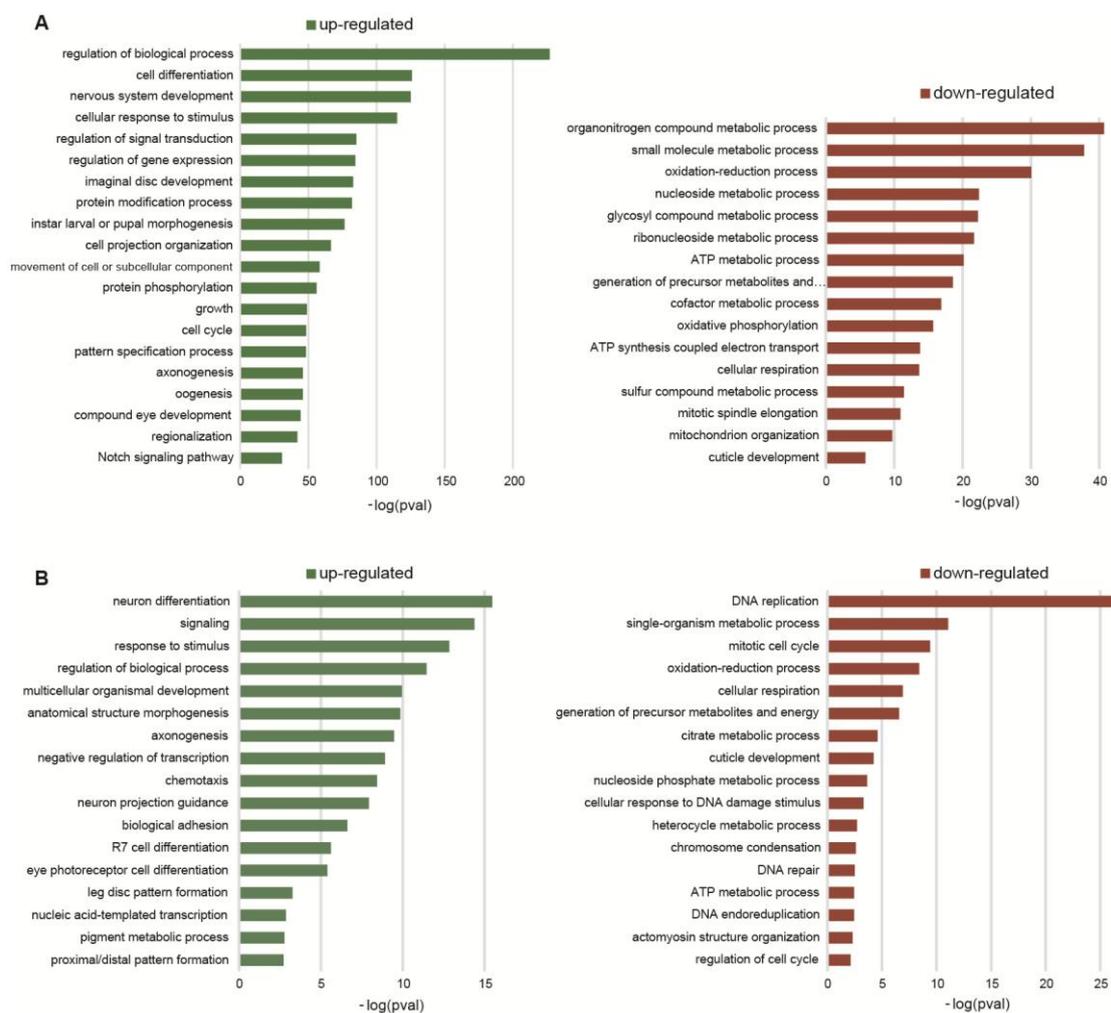


Figure 4.1.2. Biological Process GO terms enrichment. The first non-redundant enriched terms of the up- (green) and down-regulated (red) genes in the transition from **(A)** 72h to 96h transition and **(B)** 96h to 120h transition.

4.1.2 Co-expressed genes during eye-antennal imaginal disc development

In order to better characterize the different expression profiles of the genes expressed in the eye-antennal imaginal discs I performed a co-expression clustering analysis. For that I

used a recently published method that is based on Poisson Mixture models and, unlike other commonly used methods such as the k-means algorithm or hierarchical clustering, it is able to estimate the best number of clusters to describe the data (Rau et al., 2015).

First I used HTSFilter (Rau et al., 2013) to identify the genes that were expressed in at least one of the three stages (9,194 genes) and only those were introduced to HTSCluster (Rau et al., 2015), allowing to test between 1 and 25 clusters. This method outputs the results with four different model selection criteria, and the user can choose among them. These models predicted different number of clusters to describe the data: ICL = 25, BIC = 25, Djump = 13 and DDSE = 19. Previous analysis (data not shown) had shown that ICL and BIC models predicted a number of clusters always as large as the highest number of clusters it has been selected to test, so I discarded the results from these models. Inspection of the 19 clusters estimated by the DDSE model showed that some of them were redundant (data not shown), and the 13 clusters estimated by Djump were sufficient to describe all the expression profiles present in the data (Figure 4.1.3). The model reports that for a total of 8,836 genes there is a high confidence that they are placed in the correct cluster (MAP > 99%), and therefore I discarded the rest of genes from further analysis.

I ordered the 13 predicted clusters predicted according to their expression profile (Figure 4.1.3): four clusters contain clearly early expressed genes, two of them contain genes expressed only at 72h (cluster 1 and 2) and two contain genes predominantly expressed early but also with low expression at 96h and/or 120h (clusters 6 and 8); one cluster shows down-regulation at 96h but a peak of expression again at 120h (cluster 5); the largest clusters present almost constant expression throughout the three stages (clusters 12 and 10); one cluster shows constant expression at 72h and 96h and down-regulation at 120h (cluster 7); one cluster shows a peak of expression at 96h (cluster 3) and four clusters contain clearly genes with a late expression, one with high and constant expression at 96h and 120h (cluster 9), two with up-regulation in both transitions (cluster 13 and cluster 11) and one with genes expressed only at 120h (cluster 4).

The enriched GO terms for each of these ordered clusters (Table 4.1.2) describe the different events and processes that take place during the development of the larval eye-antennal imaginal discs in more detail than when using only pair-wise up- and down-regulation analysis (Figure 4.1.2).

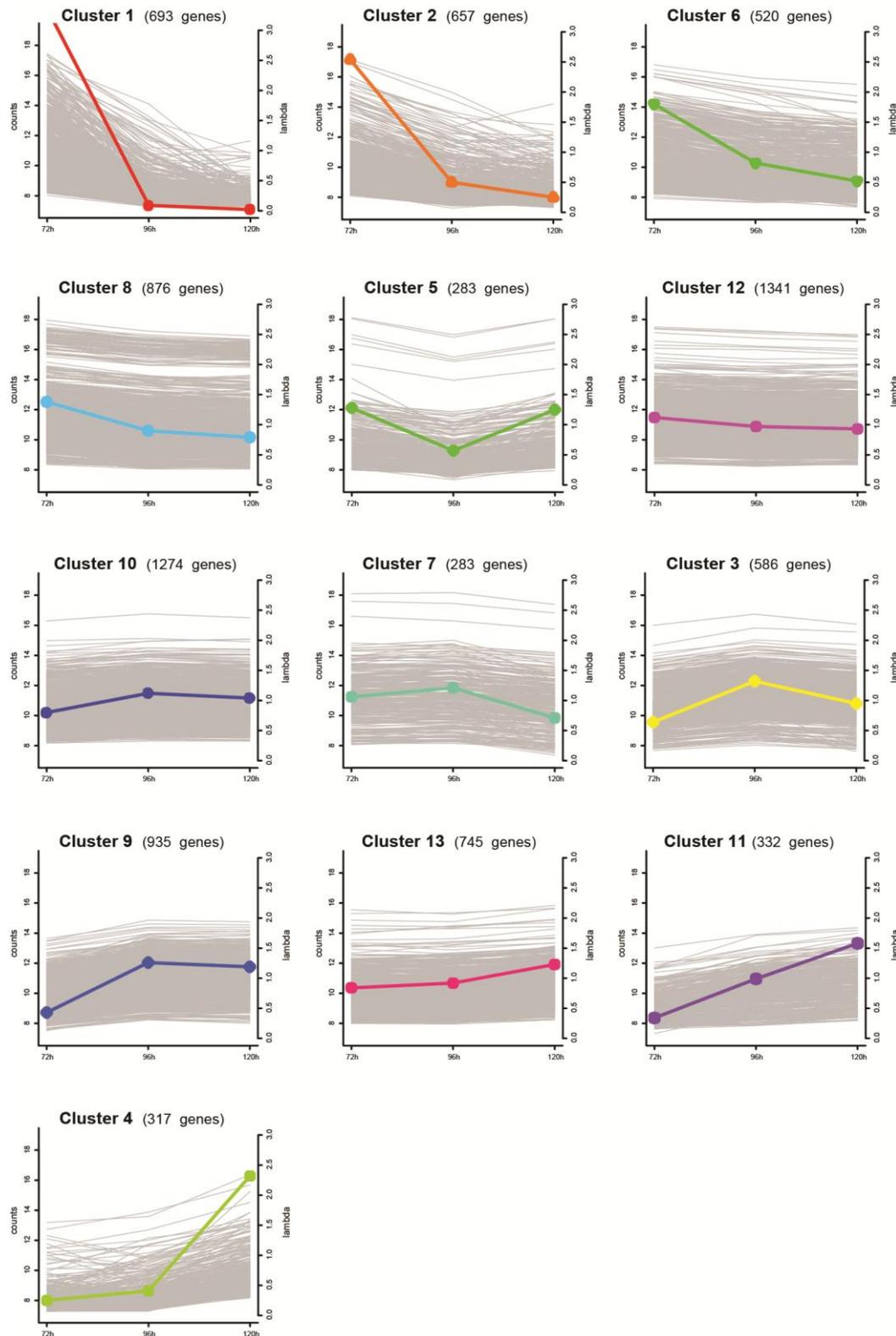


Figure 4.1.3. *D. melanogaster* expression clusters. 13 clusters predicted by HTScluster (Rau et al., 2015) (model Djump). Only genes with MAP > 99% are presented in the plots' titles and represented as background grey lines. The cluster number corresponds to the predicted output of HTScluster. The clusters are shown in the order according to expression profile of the respective genes: from expression only at 72h to expression only at 120h.

To examine if relationships between the genes belonging to each cluster were already known, I searched which published genetic interactions existed between genes of the same cluster (Table 4.1.2). It can be clearly observed that the clusters of early genes have very few known interactions among their members, while clusters of genes expressed in LIII discs have a large number of known interactions.

Table 4.1.2. GO terms of predicted clusters and genetic interactions.

Cluster	GO terms	# Genes*	# Interactions**
Cluster 1	cuticle development, aminoglycan metabolic process, body morphogenesis, humoral immune response	4	2
Cluster 2	oxidation-reduction process, single-organism metabolic process, negative regulation of peptidase activity, immune response,	9	6
Cluster 6	ATP metabolic process, electron transport chain, cellular respiration, single-organism biosynthetic process	3	2
Cluster 8	translation, gene expression, mitotic spindle elongation, ncRNA metabolic process	7	4
Cluster 5	-	0	0
Cluster 12	cellular metabolic process, cellular localization, tRNA processing, ribonucleoprotein complex biogenesis	25	16
Cluster 10	cellular macromolecule metabolic process, biological regulation, RNA processing, cellular response to stress	97	84
Cluster 7	DNA replication, cell cycle, cytoskeleton organization, neurogenesis	11	6
Cluster 3	cellular component organization, biological regulation, cell cycle, cell differentiation	106	113
Cluster 9	biological regulation, imaginal disc development, cell differentiation, generation of neurons	226	505
Cluster 13	biological regulation, transcription from RNA polymerase II promoter, RNA metabolic process, regulation of cell proliferation	29	23
Cluster 11	system development, generation of neurons, taxis, compound eye morphogenesis	77	106
Cluster 4	generation of neurons, puparial adhesion, pigment metabolic process, response to stimulus	26	18

*number of genes in the cluster that have a known interaction with at least one other gene in the cluster.

**number of genetic interactions between two genes from the cluster.

This interconnectivity can be better seen in a graphic representation (Figure 4.1.4). The early clusters (Figure 4.1.4A) have virtually no known genetic interactions with each other, although the GO term enrichment analysis shows that they clearly are involved in related biological processes and my RNA-seq data indicates that they are co-expressed during eye

and head development. In striking contrast, the genes classified in the clusters of late expressed genes (Figure 4.1.4B) have numerous known interactions between them and high interconnectivity, with the best example being cluster 9 (Figure 4.1.4C) with 505 known interactions and with at least 6 clear hubs (*Hairless (H)*, *Notch (N)*, *Epidermal growth factor receptor (Egfr)*, *Cyclin E (CycE)*, *armadillo (arm)* and *wingless (wg)*) with more than 24 genetic interactions with other genes of the cluster.

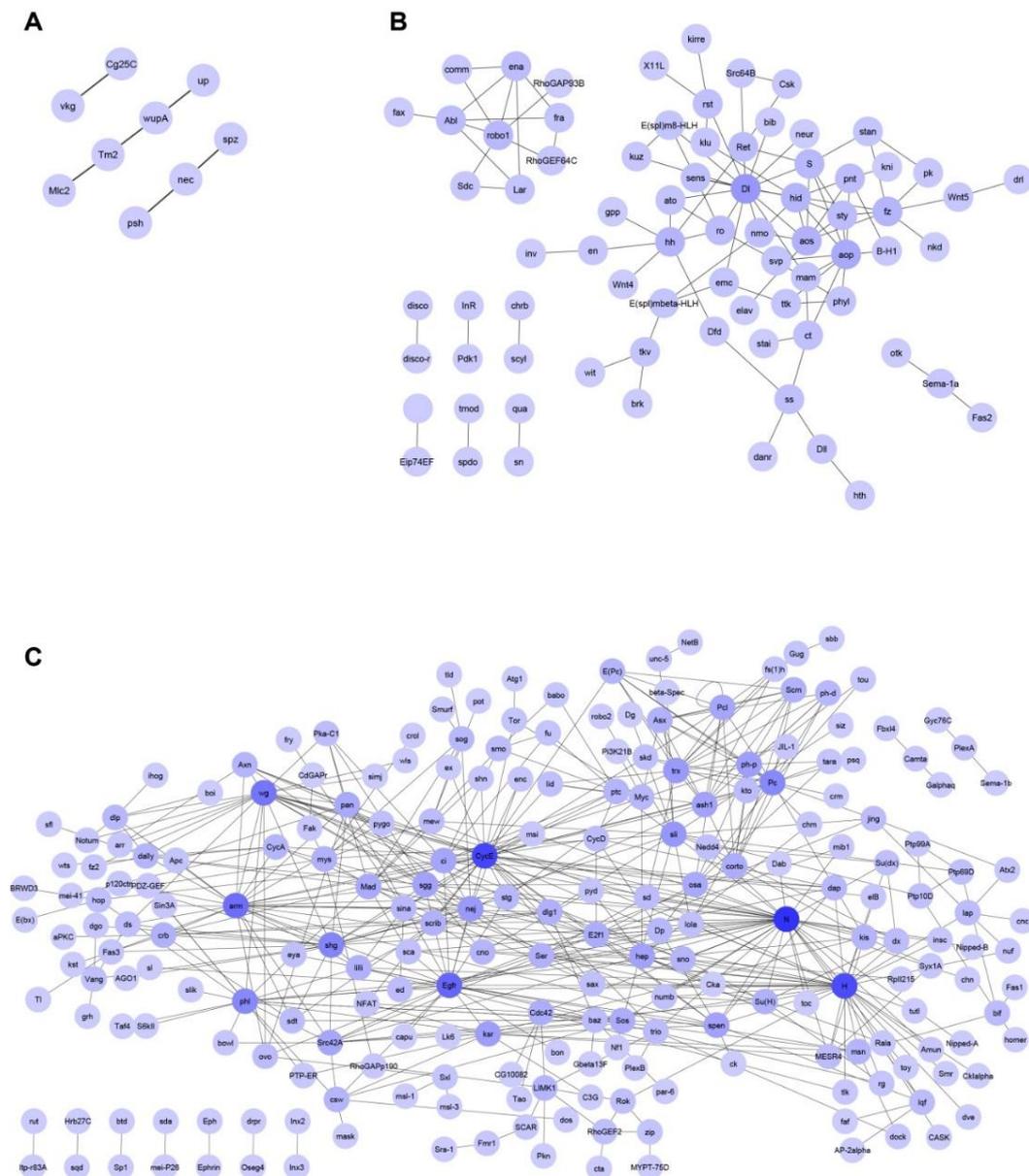


Figure 4.1.4. Networks of genetic interactions. Nodes are genes and edges are known genetic interactions. The color shading represents the number of interactions of each gene (from light blue to dark blue). **(A)** Cluster 2. Only three independent, linear interactions are known. **(B)** Cluster 11. **(C)** Cluster 9.

4.1.3 Transcription factors regulating *Drosophila* head development

One of the reasons for genes to be co-expressed could be that they are co-regulated by the same transcription factors or combinations thereof. Taking advantage of the available data of ChIP-chip and ChIP-seq experiments in *Drosophila*, I investigated what transcription factors are enriched to bind to the regulatory regions of a significant number of genes in each cluster. To do that I used the method i-cisTarget (Herrmann et al., 2012). The authors of this program have divided the non-coding regions of the genome of *D. melanogaster* into a large number of smaller regions. They have linked the presence of known *cis*-regulatory elements information coming from a large collection of experimental datasets (ChIP-seq, DNA-seq and motif discovery) to each of these regions, and then each of these regions to the adjacent coding genes, which are likely to be regulated by these *cis*-regulatory elements. When the user introduces a list of co-expressed genes, the method performs a statistical ranking to find enriched regulatory features in the regulatory regions of this set of genes. The significance of the predicted enrichment of a regulatory element for that set of genes is reported as normalized enrichment score (NES), and a higher score indicates a more significant result. In this case I used their collection of ChIP datasets, which include those published by the modENCODE Consortium (Celniker et al., 2009), by the Berkeley *Drosophila* Transcription Network Project (Li et al., 2008) and by the Furlong Lab (Zinzen et al., 2009; Junion et al., 2012).

Table 4.1.3 summarizes the results for this analysis and it lists all transcription factors that present a $NES \geq 3.0$. The transcription factor Caudal is found in the first two clusters and the co-factor Nejire is found enriched in up to 9 clusters. Clusters 8, 12, 11 and 4 are the ones with most enriched transcription factors, with more than 5 each. The clusters with genes expressed later are mostly enriched for transcription factors known to play a role in retinal development, such as Sloppy paired 1 (Slp1) (Sato and Tomlinson, 2007), Mothers against dpp (Mad) (Wiersdorff et al., 1996) or Daughterless (Da) (Lim et al., 2008).

Table 4.1.3. i-cisTarget results for each cluster.

Cluster	Total regions*	Transcription factor	Stage	Experiment	NES†	Highly ranked regions**
Cluster 1	3228	Caudal	adult female	Celniker et al., 2009	9.58	227
		Snail	emb. 4-5	Li et al., 2008	3.96	136
		Bagpipe	emb. 10-11	Zinzen et al., 2009	3.66	124
		Biniou	emb. 12-13	Zinzen et al., 2009	3.32	47
		Fushi tarazu	emb. 5	Li et al., 2008	3.03	20
Cluster 2	4596	Caudal	adult female	Celniker et al., 2009	8.12	294
		Nejire	emb. 17	Celniker et al., 2009	4.54	229

		Myocyte enhancer factor 2	emb. 12-13	Zinzen et al., 2009	3.71	65
Cluster 6	4286	Nejire	emb. 17	Celniker et al., 2009	4.46	264
Cluster 8	4891	Pannier	emb. 10-11	Junion et al., 2012	9.70	874
		dTFIIB	emb. 4-5	Li et al., 2008	6.25	672
		Nejire	larva LI	Celniker et al., 2009	5.70	526
		Medea	emb. 14	Li et al., 2008	4.99	677
		Dorsocross2	emb. 10-11	Junion et al., 2012	3.55	432
		Myocyte enhancer factor 2	emb. 10-11	Zinzen et al., 2009	3.07	486
		Pmad	emb. 10-11	Junion et al., 2012	3.01	486
Cluster 5	2400	Ecdysone receptor	prepupa	Celniker et al., 2009	5.82	154
Cluster 12	6551	Pannier	emb. 8-9	Junion et al., 2012	10.03	1269
		dTFIIB	emb. 4-5	Li et al., 2008	5.66	943
		Dorsocross2	emb. 10-11	Junion et al., 2012	5.03	650
		Pmad	emb. 10-11	Junion et al., 2012	4.51	672
		Nejire	larva LI	Celniker et al., 2009	4.50	548
		Biniou	emb. 10-11	Zinzen et al., 2009	3.54	98
		Medea	emb. 14	Li et al., 2008	3.10	799
Cluster 10	9132	Pannier	emb. 10-11	Junion et al., 2012	9.27	1119
		Pmad	emb. 10-11	Junion et al., 2012	6.97	757
		Dorsocross2	emb. 10-11	Junion et al., 2012	4.54	614
		Nejire	larva LIII	Celniker et al., 2009	3.32	540
Cluster 7	2503	Nejire	adult male	Celniker et al., 2009	6.23	222
		Pannier	emb. 10-11	Junion et al., 2012	4.30	200
		dTFIIB	emb. 4-5	Li et al., 2008	3.08	185
Cluster 3	7143	Pannier	emb. 8-9	Junion et al., 2012	5.31	510
		Nejire	larva LIII	Celniker et al., 2009	3.46	469
		Pmad	emb. 10-11	Junion et al., 2012	3.33	224
Cluster 9	15376	Nejire	larva LIII	Celniker et al., 2009	8.32	1049
		Dorsal	emb. 4-5	Li et al., 2008	4.95	1098
		Zelda	emb. 5	Li et al., 2008	3.92	976
		dTCF	emb. 10-11	Junion et al., 2012	3.54	910
Cluster 13	5563	Pannier	emb. 10-11	Junion et al., 2012	7.92	575
Cluster 11	7062	Ecdysone receptor	prepupa	Celniker et al., 2009	7.60	535
		Sloppy paired 1	emb. 8-9	Junion et al., 2012	5.13	465
		Hunchback	emb. 9	Li et al., 2008	4.41	489
		Myocyte enhancer factor 2	emb. 10-11	Zinzen et al., 2009	4.05	292
		Nejire	pupa	Celniker et al., 2009	3.96	430
		Tinman	emb. 8-9	Zinzen et al., 2009	3.50	331
		Twist	emb. 4-5	Li et al., 2008	3.49	405
		dTCF	emb. 10-11	Junion et al., 2012	3.29	289
Cluster 4	3944	Mothers against dpp	emb. 5	Li et al., 2008	4.39	163
		Snail	emb. 4-5	Li et al., 2008	4.30	189
		Runt	emb. 5	Li et al., 2008	3.61	164
		Ecdysone receptor	prepupa	Celniker et al., 2009	3.51	147
		Daughterless	emb. 5	Li et al., 2008	3.25	38
		Hunchback	emb. 4-5	Li et al., 2008	3.00	64

* number of regions from the i-cisTarget collection that are present in the list of co-expressed genes.

† normalized enrichment score.

** number of regions corresponding to the indicated transcription factor that are included in the total regions.

4.1.4 Validation of identified transcription factors

Most of the identified enriched transcription factors are known to play different roles during eye, antenna or head development in *Drosophila* (e.g. Kumar, 2004; Lim et al., 2008; Sato and Tomlinson, 2007; Sprecher and Desplan, 2008; Wiersdorff et al., 1996). However, I was intrigued by the presence of two well-studied transcription factors that appeared in more than one cluster and had not been previously related to head development: Myocyte enhancer factor 2 (Mef2) and Hunchback (Hb).

4.1.4.1 *mef2* expression in the eye-antennal imaginal disc

Mef2 is enriched in cluster 2, 8 and 11 and not expected to be expressed in eye-antennal imaginal discs. I checked its expression with a Gal4- driver line (Supplementary Figure 2) and it appears to be expressed in the most anterior part of the antennal disc, in a triangular domain between the antenna and the maxillary palp field. This region belongs to the peripodial membrane and participates in head eversion during metamorphosis (Haynie and Bryant, 1986).

4.1.4.2 *hb* expression in the eye-antennal imaginal disc

An interesting finding was the transcription factor Hb, which is enriched in clusters 4 and 11 and is the only one from these clusters (besides Mef2) that has no description of it playing a role in head or eye development.

I used available driver lines from the Vienna Tile Gal4 Library (Pfeiffer et al., 2008), more specifically lines VT038544 and VT038545 (Figure 4.1.5, Supplementary Figure 1), to investigate the possibility of *hb* being expressed in the eye-antennal imaginal discs. At late LIII stage a clear signal can be observed in two large cells at the base of the optic stalk in the posterior region of the eye disc. Line VT038543 (Supplementary Figure 1) was also tested and gave similar results, although less consistently (data not shown).

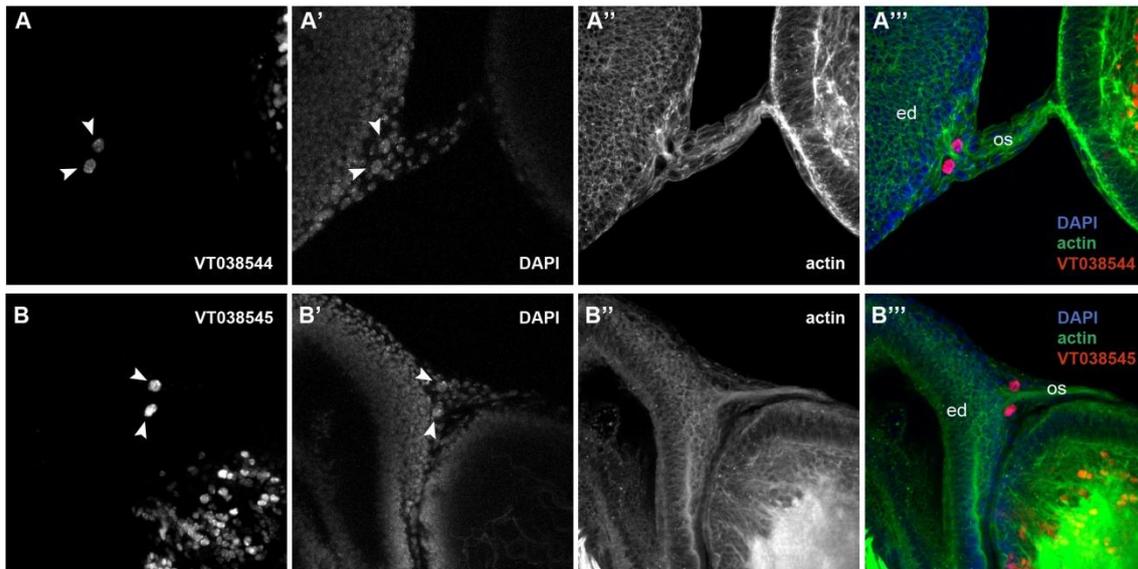


Figure 4.1.5. *hb* is expressed in the eye disc in two cells at the base of the optic stalk. Expression of histone-bound RFP driven by two adjacent genomic regions in the regulatory region of *hb* (Supplementary Figure 1). **(A)** VT038544 drives expression in two individual cells with very large nuclei (A') at the most posterior region of the eye disc (ed), at the base of the optic stalk (os). **(B)** VT038545 drives expression in the same two cells.

To confirm the expression pattern observed with Gal4- driver lines, I performed immunostaining in late LIII larval eye-antennal imaginal discs. I obtained rabbit α -Hb (kind gift from CQ Doe) and guinea-pig α -Hb (Kosman et al., 1998) antibodies and they both showed a faint signal in the same two large cell nuclei in the most posterior region of the eye disc (Figure 4.1.6 and Figure 4.1.8B, respectively). This signal was weaker than the signal in the *hb*-expressing cells in the brain (data not shown), but it was distinct in the majority of analyzed eye discs and was always found in the same region.

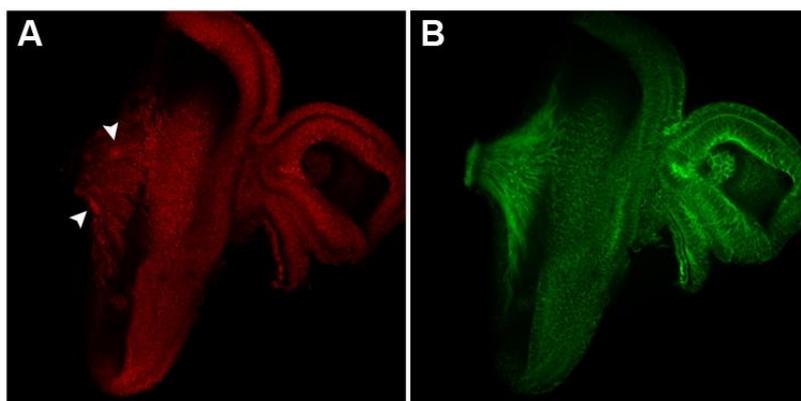


Figure 4.1.6. *hb* is expressed in the posterior region of the eye disc. **(A)** LIII eye-antennal imaginal disc stained with rabbit α -Hb antibody. Signal can be recognized in two cells on the posterior part of the disc (white arrowheads). **(B)** Phalloidin staining shows that, at this stage, the region where *hb* is expressed contains a large number of photoreceptor axons that project to the optic stalk. This picture was taken by the student Gordon Wiegleb during his Bachelor Thesis under my supervision.

4.1.5 *hb* is expressed in retinal sub-perineural glia cells

Motivated by this consistent results of *hb* being expressed in the eye imaginal disc I decided to further investigate its possible new role in visual system development. A better observation of the Z-stack projections of the immunostaining and driver lines expression revealed that the expression in the disc proper was basal to the differentiating photoreceptors (data not shown). It has previously been shown that various glia cell types are present in this part of the eye-antennal imaginal disc, supporting the developing photoreceptors (Choi and Benzer, 1994). These retinal glia cell types include migratory surface glia (including perineural and sub-perineural glia cells) and wrapping glia. During LIII stage, surface glia cells enter the eye disc through the optic stalk and migrate towards the anterior part of the disc, remaining always posterior to the advancing morphogenetic furrow (Choi and Benzer, 1994; Rangarajan et al., 1999). When photoreceptors differentiate, the contact of their growing axons with perineural glia cells triggers the reprogramming of these glia cells into differentiated wrapping glia (Silies et al., 2007). These glia cells extend their cell membrane to ensheath bundles of axons that project to the brain lobes through the optic stalk (Hummel et al., 2002). Perineural glia cells are in the most basal part of the disc and above them are two sub-perineural glia, known as carpet cells, that separate them from the projecting axons and the differentiated wrapping glia cells (Figure 4.1.7). Carpet cells have polyploid nuclei and strikingly large cell bodies, each of them covering half of the retinal field (Silies et al., 2007).

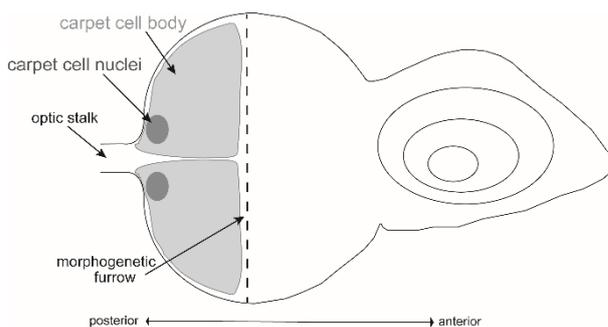


Figure 4.1.7. Schematic representation of the carpet glia cells on the eye imaginal disc. Two large glia cells known as carpet cells are present on the eye region (light grey area), behind the morphogenetic furrow. These cells have very large, polyploid nuclei (dark grey circles). These cells serve as surface for other glia cells to facilitate and coordinate their migration into the eye disc to find the nascent photoreceptor axons. Figure adapted from Silies et al. 2007.

Co-staining with the pan-glial marker Repo (Figure 4.1.8A) and the sub-perineural glia marker Moody (Schwabe et al., 2005) (Figure 4.1.8B) showed that the two cells expressing *hb* are sub-perineural glia cells. Their large nucleus size and the position at the posterior region of the eye disc also confirm this fact.

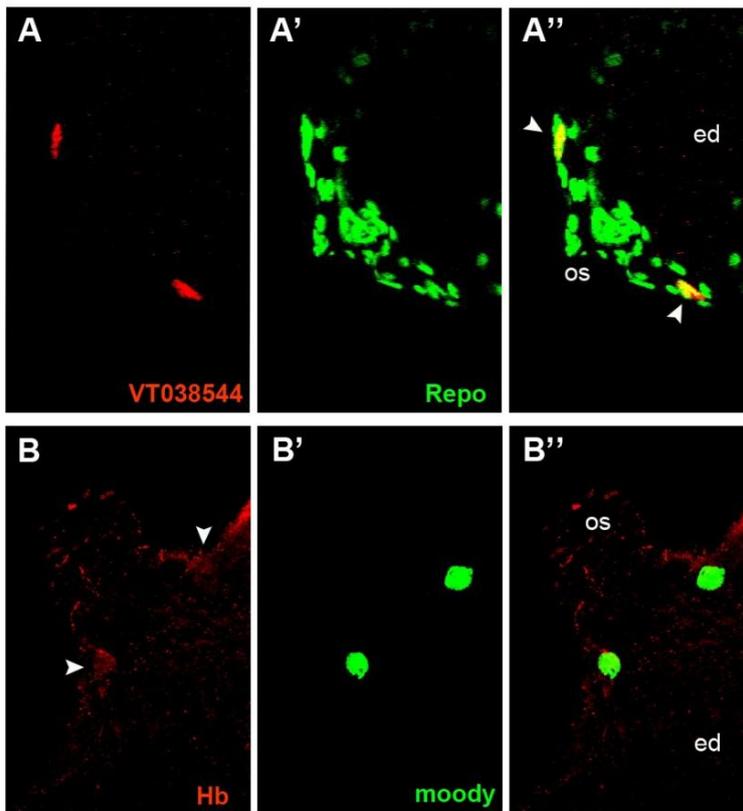


Figure 4.1.8. *hb* is expressed in sub-perineural glia cells. Co-staining of Hb with other glia markers. “ed”: eye disc; “os”: optic stalk. **(A)** Co-staining of *hb* (as visualized with a *hb*-Gal4 driver line crossed with UAS-H2B-RFP reporter) and rabbit α -Repo antibody. The *hb*-expressing cells also express *repo* (arrowheads), indicating that they are glia cells. **(B)** Co-staining of rabbit α -Hb antibody and *moody* (*moody*-Gal4 driving UAS-GFP expression). The two cells expressing *hb* are also *moody*-positive (arrowheads), indicating that they are sub-perineural glia cells.

One important feature of carpet cells is that they migrate through the optic stalk into the eye-antennal imaginal disc (Choi and Benzer, 1994; Silies et al., 2007). In order to test whether the *hb* positive cells also show this behavior, I followed the expression of the *hb* driver lines at LII and LIII larval stages (Figure 4.1.9). I could corroborate that these cells indeed migrate through the optic stalk during LII and early LIII stage, and then enter the disc and remain at the posterior region of the disc, adjacent to the optic stalk. Already by LII stage their cell nuclei can be easily recognized by their large size (Figure 4.1.9A). By late LIII they sit at the right and left sides of the optic stalk inside the eye disc (Figure 4.1.9C), and they are never found in the midline of the retinal field (Silies et al., 2007).

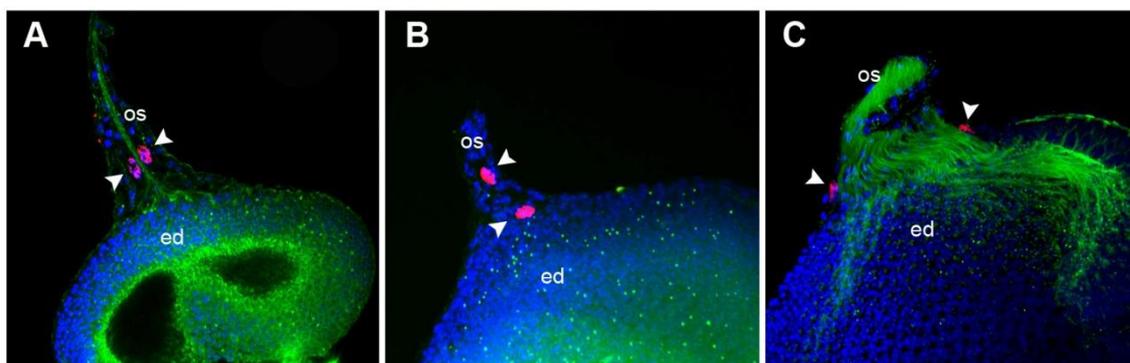


Figure 4.1.9. Cells expressing *hb* migrate through the optic stalk into the disc during larval stages. **(A-C)** *hb*-expressing cells are visualized with VT038544 (*hb*-Gal4) driving histone-bound RFP (red), actin is shown by Phalloidin staining (green) and the cell nuclei with DAPI (blue). “ed”: eye disc; “os”: optic stalk. **(A)** Eye disc at LII stage (72h AEL). The cells expressing *hb* are in the

optic stalk and their bigger nuclei can be already recognized. **(B)** Eye disc at mid LIII stage (96h AEL). Carpet cells are entering the eye disc. **(C)** Eye disc at late LIII stage (120h AEL). The *hb*-expressing cells are in the eye disc margin, each at one side of the optic stalk base. These pictures were taken by the student Julia Schneider during her Master's Lab Rotation under my supervision.

4.1.5.1 *hb* is not expressed in brain sub-perineural glia

To investigate if *hb* is expressed in other glia cells, I checked if it is also co-expressed with the pan-glial marker *Repo* in the brain (Figure 4.1.10). In brains of both LII and LIII stage I was able to identify only one cell showing overlapping expression of *hb* and *repo*, although it is not clear if it is the same cell in the two stages. This cell(s) is located on the right side of the brain, at the edge of the optic lobe and near the central brain. No *hb* expression could be detected in sub-perineural glia cells, which are located on the brain surface and can be identified by *moody* expression (data not shown).

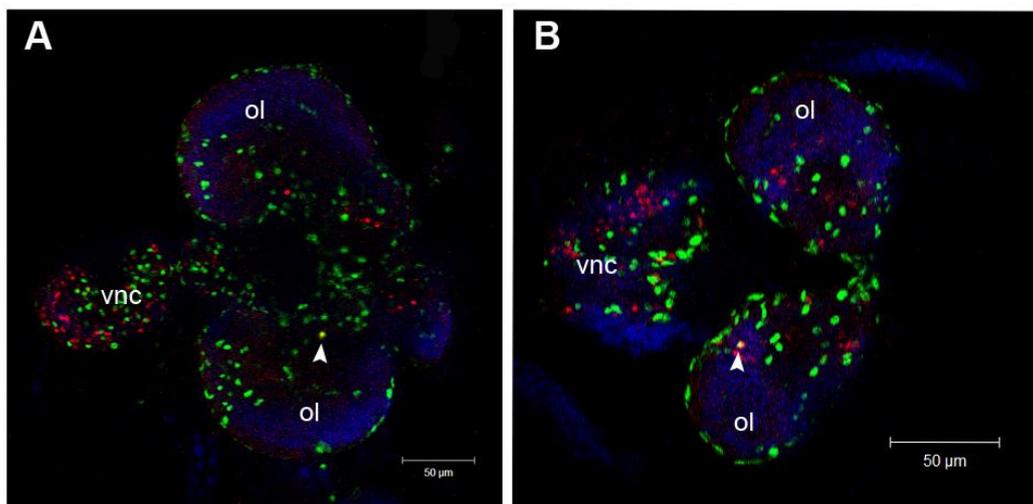


Figure 4.1.10. *hb* and *repo* expression in the brain. **(A-B)** guinea-pig α -Hb antibody in red, *repo*>>GFP in green. “ol”: optic lobe; “vnc”: ventral nerve cord. **(A)** Larval brain at LII stage. Only one cell (arrowhead) on the right side of the brain can be observed that is positive for both *hb* and *repo*, near the margin between the optic lobe and the central brain. **(B)** At late LIII stage, also only one glia cell could be identified that might express *repo*. This cell is in the right optic lobe and more posterior than the cell at LII (although it could be the same cell that has migrated).

4.1.6 Hb function in the development of retinal glia

To better understand the possible role of Hb during carpet cell development I performed different loss and gain of function experiments using available glia cell driver lines. Since sub-perineural glia cells form the blood-brain barrier (Carlson et al., 2000), a possible effect of the loss of Hb in the integrity of this structure was also analyzed.

4.1.6.1 Hb loss of function experiments

To study the effects of Hb loss of function in the carpet cells, first I used an RNAi approach. I obtained 4 different UAS-*hb*_{dsRNA} lines from Bloomington Stock Center (#54478, #29630 and #34704) and from the Vienna Drosophila Research Center (#107740). In order to evaluate the knock-down efficiency I took advantage of the fact that Hb is known to be necessary during early embryogenesis (Lehmann and Nüsslein-Volhard, 1987; Nüsslein-Volhard and Wieschaus, 1980). I crossed the UAS-*hb*_{dsRNA} flies with the *hb*-Gal4 lines (VT038544 and VT038545) to see if this indeed affected the survival of the offspring. Only one of the RNAi lines, namely #34704, produced no adult flies and few dead pupae when crossed with the *hb*-Gal4 flies. The other three lines produced a normal number of offspring with no obvious phenotype. Consequently, I used the #34704 line for the following knock-down experiments. Please note that the evaluation of knock-down efficiency in the developing eye-antennal imaginal discs using quantitative PCR is very limited due to the fact that the expression of *hb* is very low (practically no reads are detected by RNA-seq, not shown).

I also used a temperature sensitive *hb* mutant (*hb^{ts}*) (Bender et al., 1987) to investigate the effects of loss of Hb function. Since Hb is necessary during embryogenesis, the analyzed flies were kept at 18°C during egg collection procedure, and only transferred to the restrictive temperature of 28°C during larval stages (either at LI or LII stage).

LIII eye-antennal imaginal discs were analyzed from knock-down (*repo*>>*hb*_{dsRNA} and *moody*>>*hb*_{dsRNA}) and mutant (*hb^{ts}*) flies and diverse phenotypes were observed. The most common was the absence of one or both of the large carpet cell nuclei (Figure 4.1.11). Carpet cell nuclei were easily identified by α -Repo staining because of their large size and their position on the posterior end of the eye disc on each side of the base of the optic stalk (Figure 4.1.11A). In wild type animals the two carpet cells could be observed almost in 90% of the eye-discs. In most cases where only one carpet cell nucleus could be identified it was due to technical problems like folding of the disc during mounting or because rests of the optic lobes covered the retinal field region. In contrast, in 30% to 38% of the studied Hb loss of function discs only one carpet cell nucleus was observed in the eye discs (Figure 4.1.11B and D). In some cases, this single polyploid Repo-positive nucleus was located in the midline of the retinal field (Figure 4.1.11B). In other discs no carpet cell nuclei could be observed in the retinal field of the eye-antennal imaginal discs (Figure 4.1.11C and D). The number of discs with no observable carpet cell nuclei varied greatly according to the experiment: the *repo* driven *hb* RNAi resulted in 12% of discs without

carpet cell nuclei and in the *hb^{ts}* flies that were transferred to the restrictive temperature at LI about 38% of the discs did not possess carpet cell nuclei (Figure 4.1.11D). A slightly larger percentage of discs with no carpet cell nuclei were observed when the larvae were transferred to the restrictive temperature at LI stage in comparison to when they were transferred to 28°C during LII stage.

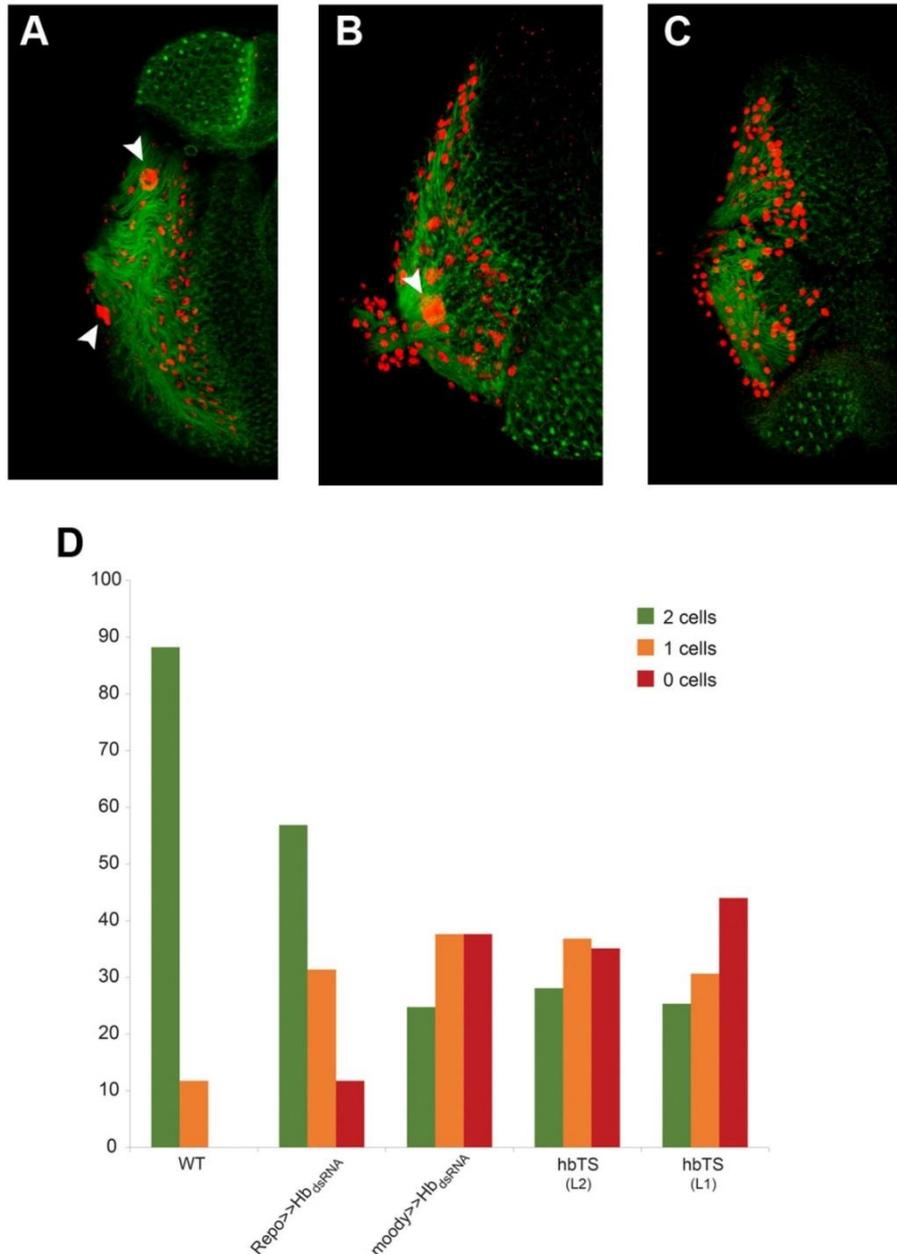


Figure 4.1.11. Hb loss of function results in loss of carpet cell nuclei. (A-C) Late LIII eye discs stained with rabbit α -Repo (red) and Phalloidin (green). (A) In wild type discs, carpet cells can easily be recognized by their large nuclei stained with Repo (arrowheads). (B) A phenotype observed in knock-down experiments (in this picture, *repo>>hb^{dsRNA}*), is the presence of only one carpet cell as observed by Repo staining (arrowhead). In some cases, this cell can be in the midline of the retinal field, which is never the case in wild type. (C) Another phenotype observed as a result of Hb loss of function experiments is the absence of carpet cell nuclei. In this picture, *hb^{ts}* flies moved to restrictive temperature during LII stage. (D) Quantification of observed phenotypes in

Hb loss of function experiments as count of observable carpet cells by Repo staining. (wild type $n=34$, $repo >> bb_{dsRNA}$ $n=51$, $moody >> bb_{dsRNA}$ $n=101$, bb^{ts} (LI) $n=57$, bb^{ts} (LII) $n=75$). In wild type two carpet cells can be observed in 90% of the discs, and at least one can always be identified. *bb* RNAi driven by *repo* reduces the percentage of discs with 2 carpet cells, and in some cases no carpet cell can be identified. The phenotypes in *moody* driven *bb* RNAi and *bb^{ts}* mutants are even stronger, with the strongest phenotype observed when *bb^{ts}* mutant larvae are moved to the restrictive temperature at LI stage.

Besides the loss of carpet cell nuclei, other phenotypes were observed. In parts of the retinal field where also carpet cells were missing I observed absence of other glia cells (compare Figure 4.1.12A' to B'). Co-staining with HRP was used to visualize axon projections. This analysis revealed that axon projections were in some cases unorganized (compare Figure 4.1.12A to B). And eventually, for some LII stage discs, I observed that glia cells prematurely migrated into the eye disc before photoreceptor differentiation started (data not shown).

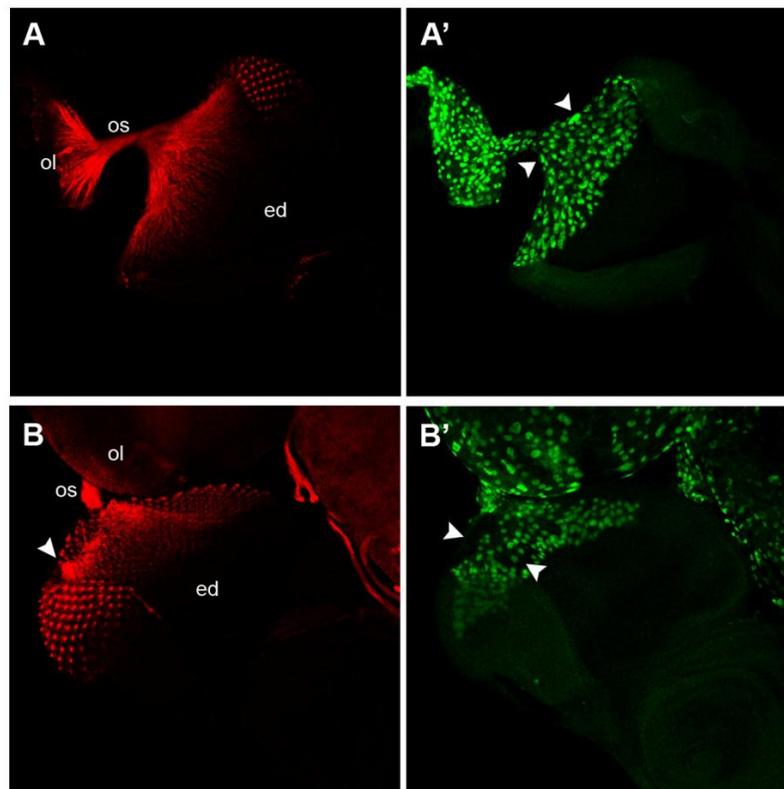


Figure 4.1.12. Hb loss of function affects axon projection and the organization of other retinal glia cells. Late LIII eye imaginal discs attached to the optic lobe immunostained with HRP (red) and Repo (green). “ol”: optic lobe; “os”: optic stalk; “ed”: eye disc. **(A)** In wild type larvae, axons project in an organized manner from the developing photoreceptors in the eye disc into the optic lobes through the optic stalk (red). **(A')** Glia cells occupy all the basal surface of the eye disc posterior to the morphogenetic furrow to support the developing photoreceptors and their axons. Carpet cells can be observed at the posterior margin of the eye disc (arrowheads). **(B)** In some $repo >> bb_{dsRNAi}$ larvae, axons don't project correctly and form unorganized bundles (arrowhead). **(B')** In the basal surface of the eye disc some patches without glia cells can be observed (arrowheads), and carpet cells cannot be identified.

4.1.6.2 Hb gain of function experiments

To investigate the role of Hb in retinal glia cells, I also performed overexpression analyses in glia cell types other than the sub-perineural glia. Misexpression of *hb* in all glia cells (*repo>>hb*) prevented embryos to hatch and therefore no animals could be analysed. When *hb* was overexpressed in larval perineural glia cells (*c527-Gal4* (Ito et al., 1995) was crossed to *UAS-hb*), most larvae died before they reached LIII stage. Only a few larvae at LII stage could be studied. In these animals, retinal glia cells in the optic stalk seemed to be bigger and the carpet cells were not recognizable among them, which usually are at this stage. It also seemed like these larvae could have more glia in the optic stalk than expected at LII stage, although no thorough quantification could be performed. A more detailed analysis is needed since very few larvae could be studied.

I also performed *hb* overexpression in wrapping glia (*Mz97-Gal4* (Ito et al., 1995) was crossed to *UAS-hb*) and larvae could develop normally until pupal stage. A closer look at the eye-antennal imaginal discs showed that in these animals, glia cell nuclei were located between the axon bundles in the optic stalk, which was never the case in wild type animals (Figure 4.1.13). The cell bodies and nuclei of perineural glia in the optic stalk form a single outer layer surrounding the axon bundle (Figure 4.1.13A), and these cells were never located between the axon bundles. In contrast, wrapping glia cell nuclei remain on the eye disc or at the most anterior part of the optic stalk, and only their cell bodies project together with the photoreceptor axons bundles through the optic stalk into the lamina (Hummel et al., 2002; Murakami et al., 2007; data not shown). When wrapping glia overexpress *hb*, glia cell nuclei are misplaced inside the axon bundles in the optic stalk (Figure 4.1.13B).

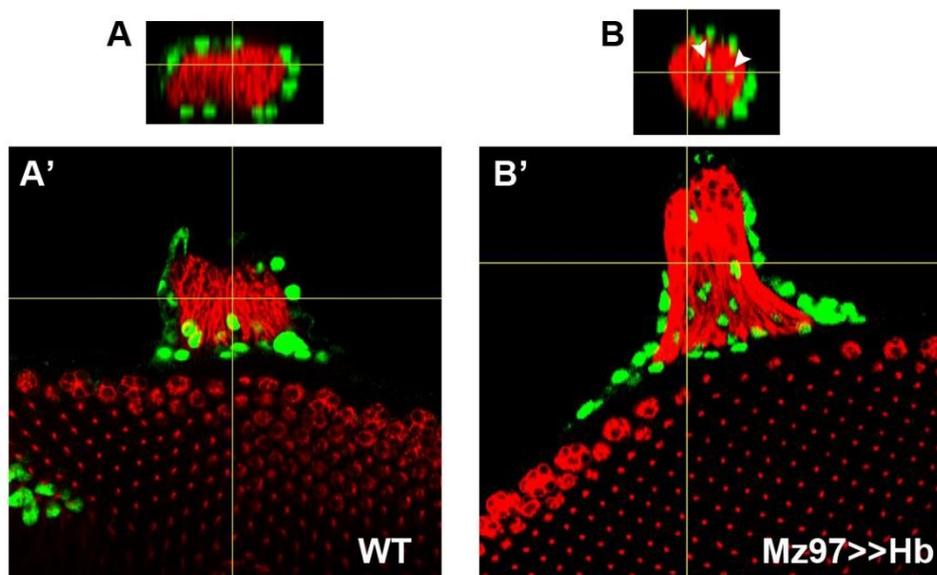


Figure 4.1.13. *hb* overexpression in wrapping glia. Immunostaining with Repo (green) and HRP (red). **(A and B)** z-section of the optic stalk corresponding to the horizontal yellow line in **A'** and **B'**. **(A)** Wild type 120h eye disc. In the optic stalk, the photoreceptor axons are organized in a single bundle and ensheated by a monolayer of glia cells. Glia cell nuclei are never found inside this axon bundle in wild type (Hummel et al., 2002). **(B)** 120h eye disc where wrapping glia cells (*Mz97*) are overexpressing *hb*. Glia cells, as recognized by Repo (green) staining, can be observed inside the optic stalk and completely surrounded by axonal projections.

4.1.6.3 Loss of Hb function results in blood-brain barrier defects

Sub-perineural glia cells (including carpet cells), together with the perineural glia, cover the entire surface of the brain from larval stages onwards (Figure 4.1.14A), contributing to the establishment of a protective blood-brain barrier by establishing inter-cellular septate junction (Carlson et al., 2000). The blood-brain barrier prevents the substances that circulate in the hemolymph to enter the brain and helps maintaining the proper homeostatic conditions of the nervous system (Edwards et al., 1993). I decided to investigate if the loss of Hb in developing carpet cells had an effect on the integrity of this blood-brain barrier, since our previous analyses indicated that loss of Hb function interferes with carpet cell formation. To do this I injected fluorescently labeled dextran into the abdomen of *moody>>hb_{dsRNA}* adult flies (Figure 4.1.14B). Nearly all wild type animals with a properly formed blood-brain barrier presented a fluorescent signal in their body but not in the brain nor in the retina (Figure 4.1.14C). However, in animals that have an incomplete blood-eye barrier, the dextran penetrated into the retina and fluorescence was observed in the compound eyes (Figure 4.1.14D). Since it is known that blood-brain barrier permeability can increase after exposure to stress conditions (Sharma and Dey, 1986; Skultétyová et al., 1998), I only scored animals that survived 24h after the injection of dextran. In most cases, the two eyes of an individual presented different fluorescent intensity, and even no fluorescence in one eye but strong signal in the other. Therefore, I scored each eye separately. Interestingly, *moody>>hb_{dsRNA}* flies had a significantly higher rate of fluorescent retinas ($p = 8.08e-7$, χ^2 test), indicating that their eyes were not properly isolated from the blood circulating in the body cavity (Figure 4.1.14E).

In summary, the loss of Hb affects the sub-perineural glia cells, either by reducing their nucleus size, by affecting their polyploidy or by affecting their presence in the eye disc all together. This has an effect on the integrity of the blood-eye barrier of adult flies.

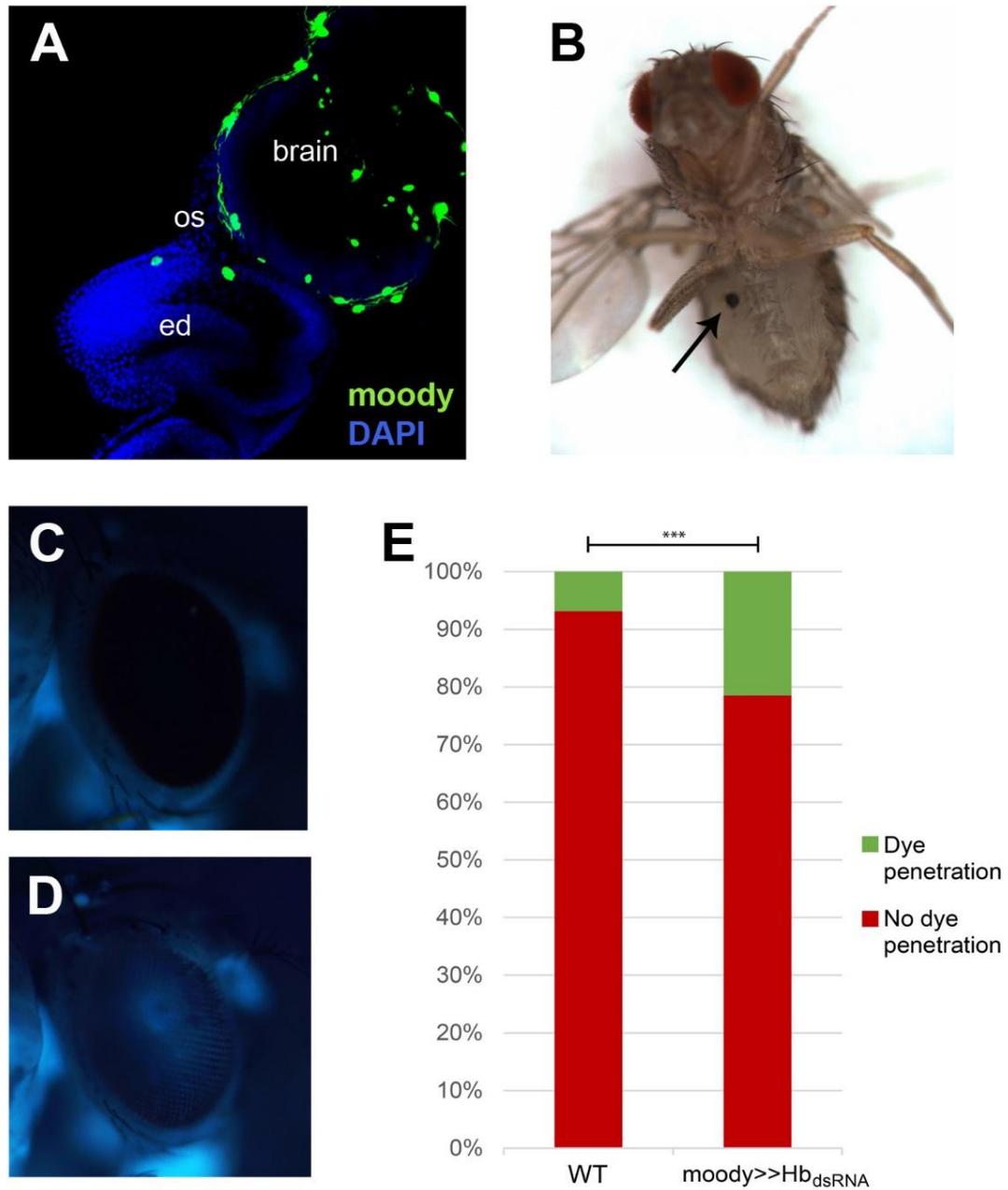


Figure 4.1.14. Blood-eye barrier integrity. (A) Sub-perineural glia cells (*moody*-positive, here stained in green) cover the brain surface with their large cell membranes and form the blood-brain barrier. Carpet cells are also sub-perineural glia cells and, in the adult, they form the blood-eye barrier. (B) To assay the integrity of the blood-eye barrier I injected fluorescent dye in the abdomen of adult flies (black arrow) and scored the presence of dye in the fly eye. (C) In flies that have a correctly formed blood-eye barrier, fluorescence can be observed in the body but not in the eye. (D) In flies with impaired blood-eye barrier, fluorescent dye can be observed in the fly eye. (E) Quantification of eyes with or without dye penetration in the eye. Knock-out flies have a significant increase in the penetrance of dye in the eye, indicating a defective blood-eye barrier. (wild type $n=262$, *moody>>hb_{dsRNA}* $n=326$). p -value = $8.08e-7$ (χ^2 test).

4.1.7 Expression of putative Hb target genes in the eye-antennal imaginal disc

Since I have detected Hb because of its target genes, I also investigated whether some of these targets are expressed in the carpet cells. The *i-cisTarget* method (Herrmann et al., 2012) to detect transcription factor enrichment in the regulatory regions of co-regulated genes is based on the arbitrary partition of the *Drosophila* genome in more than 13,000 regions. All genes included in a particular region are associated to the transcription factor binding interval, resulting maybe in an unspecific association between transcription factor and target genes. Therefore, I decided to generate a more confident list of putative Hb target genes in the eye-antennal imaginal disc. I selected only one gene for each ChIP genomic interval, i.e. the closest gene to the peak. I searched then the region around the transcription start site of each gene (1,000 bp up- and downstream) and only kept those genes that contained at least one instance of the Hb binding motif. This resulted in 1,288 putative Hb target genes. I combined this with my developmental transcriptomics data and I found that 77 of these genes are present in the clusters that have enrichment for Hb (cluster 4 and 11, Figure 4.1.15A). I searched the GO terms for biological function known for these 77 genes (Figure 4.1.15B) and found that 17 code for transcription factors and up to 25 code for proteins integral to membrane (Supplementary Table 1). A number of GO terms are related to neuronal development and eye development and to note is the presence of genes known to be related to glia cell migration and endoreduplication.

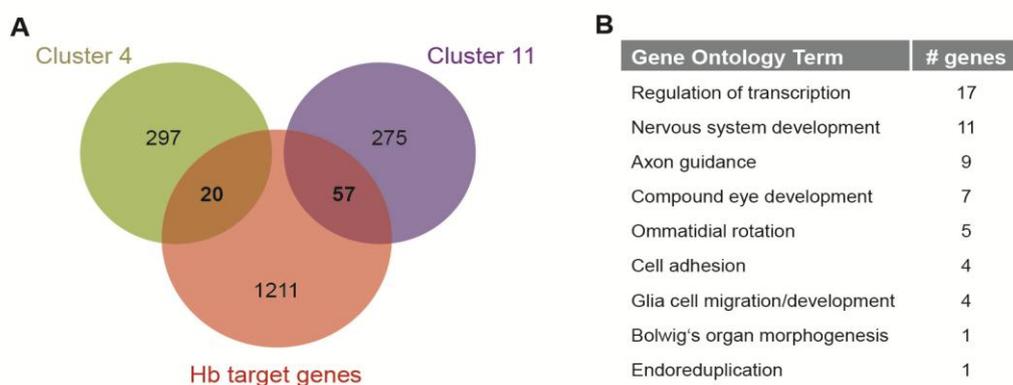


Figure 4.1.15. Hb target genes. (A) Overlap of genes in clusters 4 and 11 and high confidence Hb target genes. **(B)** Some GO terms annotated for the 77 Hb target genes in clusters 4 and 11. Full list of genes can be found in Supplementary Table 1.

Based on their annotated GO terms, predicted or known cellular location and the availability of driver lines, I selected 9 of these targets: *brinker* (*brk*), *Cadherin-N* (*CadN*), *Delta* (*Dl*), *Fasciclin 2* (*Fas2*), *knirps* (*kni*), *rhomboid* (*rbo*), *roundabout 3* (*robo3*), *Sox21b* and *Src*

oncogene at 64B (Src64B). Analysis of late LIII eye-antennal imaginal discs showed that four of these targets (*Fas2*, *rho*, *Sox21b* and *Src64B*) are expressed in the most posterior region of the disc (Figure 4.1.16), *brk* is expressed ubiquitously (Supplementary Figure 3) and *CadN* showed expression in one cell located near where the ventral carpet cell is located (Supplementary Figure 3). The lines tested for *Delta* and *kni* showed no expression in the eye-antennal imaginal disc at late LIII stage, and *robo3* only in a ventral domain in the antennal disc (data not shown).

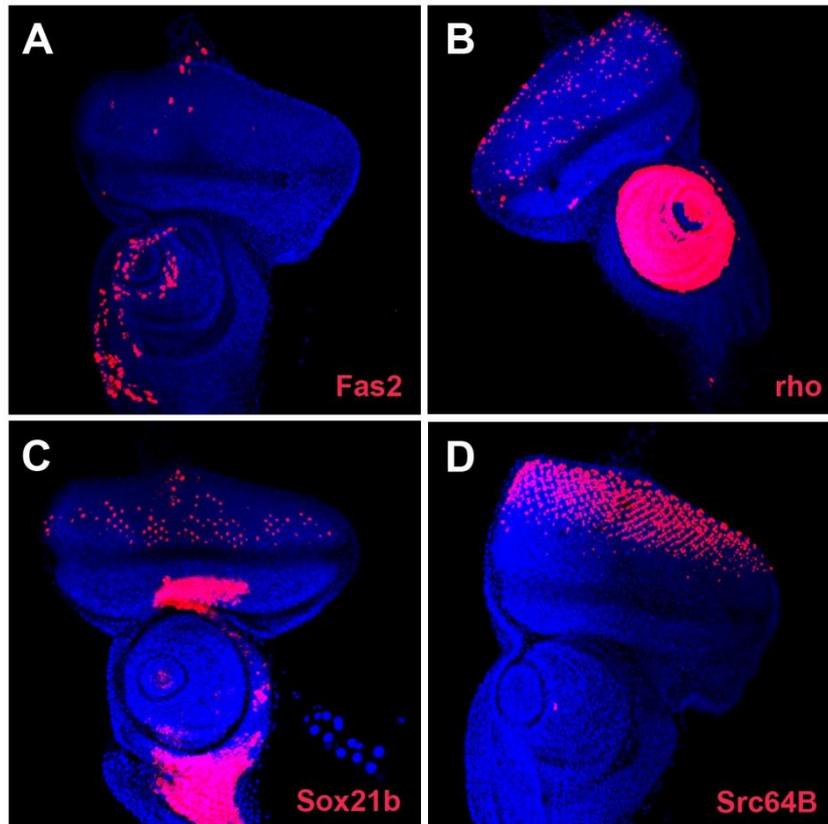


Figure 4.1.16. Expression of Hb target genes in eye-antennal imaginal discs. Gal4 driver lines crossed with UAS-H2B-RFP. All discs are from 120h AEL larvae. DAPI staining in blue. **(A)** *Fas2* is expressed in a few cells in the retinal field and optic stalk and in a long domain in the ventral side of the antennal disc. **(B)** *rho* is expressed in several cells in the retinal field (posterior to the morphogenetic furrow) and in the complete antennal domain. **(C)** *Sox21b* is expressed in regularly spaced cells posterior to the morphogenetic furrow and in two domains posterior and anterior to the developing antenna domain. **(D)** *Src64B* is expressed in photoreceptor cells that are more advanced in their development (posterior).

4.2 A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species

The following parts of my Thesis (“**Gene expression divergence in closely related *Drosophila* species**” and “**Eye size variation in two closely related *Drosophila* species**”) include the comparison of RNA-seq data from different species (*D. simulans*, *D. mauritiana* and *D. melanogaster* in the former and only between *D. simulans* and *D. mauritiana* in the later). In order to do that in an unbiased manner, I developed a method to reciprocally re-annotate the available genomes of these species. What follows is the manuscript that I have written to describe this methodology. I wrote the full first draft version of the manuscript and I and my supervisor Dr. Nico Posnien have written this final version. Dr. Isabel Almudi and Dr. Alistair P. McGregor (Oxford Brookes University, UK) are co-authors in this work. Dr. Isabel Almudi collaborated in the dissection of eye-antennal imaginal discs that were sequenced by RNA-seq (see Materials and Methods). Dr Alistair P. McGregor was involved in the initial design of the RNA-seq experiment. This manuscript is currently in revision in *BMC Genomics* (minor revision).

4.2.1 Abstract

Background

RNA-seq-based on short reads generated by next generation sequencing technologies has become the main approach to study differential gene expression. Until now, the main applications of this technique have been to study the variation of gene expression in a whole organism, tissue or cell type under different conditions or at different developmental stages. However, RNA-seq also has a great potential to be used in evolutionary studies to investigate gene expression divergence in closely related species.

Results

We show that the published genomes and annotations of the three closely related *Drosophila* species *D. melanogaster*, *D. simulans* and *D. mauritiana* have limitations for inter-specific gene expression studies. This is due to missing gene models in at least one of the genome annotations, unclear orthology assignments and significant gene length differences in the different species. A comprehensive evaluation of four statistical frameworks (DESeq2, DESeq2 with length correction, RPKM-limma and RPKM-voom-limma) shows that none of these methods sufficiently accounts for inter-specific gene length differences, which inevitably results in false positive candidate genes. We propose that published reference genomes should be re-annotated before using them as references for RNA-seq experiments to include as many genes as possible and to account for a potential length bias. We present a straightforward reciprocal re-annotation pipeline that allows to reliably compare the expression for nearly all genes annotated in *D. melanogaster*.

Conclusions

We conclude that our reciprocal re-annotation of previously published genomes facilitates the analysis of significantly more genes in an inter-specific differential gene expression study. We propose that the established pipeline can easily be applied to re-annotate other genomes of closely related animals and plants to improve comparative expression analyses.

4.2.2 Background

Comparative studies of gene expression have been used to understand the regulation of a wide range of biological processes. With the development of next generation sequencing (NGS) technologies, and in particular the use of Illumina sequencing platforms, reliable genome wide comparison of gene expression between different biological conditions has become possible (Garber et al., 2011; Ozsolak and Milos, 2011; Wang et al., 2009). Moreover, a growing number of available genome and transcriptome sequences (Ellegren, 2014; Evans et al., 2013; Haussler et al., 2009; Koepfli et al., 2015; Poelchau et al., 2014) now provides the opportunity to compare gene expression not only in well-established, but also in emerging model systems. Especially, the comparison of gene expression between both closely (Coolon et al., 2014; Graze et al., 2009, 2012; McManus et al., 2010; Paris et al., 2013; Wittkopp et al., 2004, 2008; Zhao et al., 2015) and distantly related species (Aubry et al., 2014; Brawand et al., 2011; Gerstein et al., 2014; Perry et al., 2012) has great potential to help understand phenotypic divergence and species adaptations at a mechanistic level (Musser and Wagner, 2015).

Experiments to study differential gene expression using NGS technologies (RNA-seq) are based on a sequencing library generated from reverse transcribed messenger RNA (mRNA) that is extracted from the tissue and conditions of interest. Illumina sequencing, for example, results in the generation of millions of short reads ranging from 36 bp to 150 bp (Metzker, 2010; Shendure and Ji, 2008). The first step of the bioinformatics analysis is to align these reads to a reference that represents all transcripts that should be quantified (Bray et al., 2015; Langmead et al., 2009; Li and Durbin, 2009; Trapnell et al., 2009, 2010). This reference can be a whole genome sequence with annotated gene models or a transcriptome. The latter can either be generated by a *de novo* assembly of the RNA-seq reads (Haas et al., 2013; Schulz et al., 2012) or it could be extracted from an annotated genome. The next step is to determine the number of reads that are aligned to a gene model or transcript. Depending on the type of reference used (genome or transcriptome) various different methods have been established (Anders et al., 2014; Li et al., 2009). Finally, the number of reads assigned to a given gene model or transcript is compared between different conditions to identify differentially expressed genes.

The steps outlined above for a general RNA-seq experiment are suitable to compare gene expression levels between different conditions, stages or tissues of the same species. However, comparison of gene expression between different species or populations of the same species needs to account for differences in gene sequences. In this case, reads should

be mapped to species-specific references for which the expression level of a gene in one of the species is compared to the expression level of its ortholog in the other species. Most importantly, this requires sets of orthologous genes reliably identified in all references. Since genomes or transcriptomes are usually generated by different research groups for different applications and using different pipelines for assembly and annotation, annotated references for inter-specific gene expression studies are often not comparable. For instance, orthologous genes might be missing from one or more of the references as result of natural variation or technical problems like incomplete assemblies or too many sequencing errors, which hampers unequivocal identification of orthologous genes. Additionally, it is common practice to filter out genes that are incomplete or lack synteny in relation to a model reference from new gene model predictions (Yandell and Ence, 2012). Even though there are many tools available to perform genome annotation, a general standard does not exist. Therefore, the final gene set generated by each genome project will have genes missing as a result of methodological problems and filtering criteria, and this can directly influence the result of the differential gene expression analysis (Zhao and Zhang, 2015).

Additionally, even if most one-to-one orthologs have been successfully identified in different references, these gene models may vary in length for various reasons: First, the genes could naturally differ in length among species. Second, as a consequence of differences in the sequence or assembly quality of the reference genomes (e.g. stretches of Ns or premature stop codons due to sequencing errors, incorrect scaffolding or repetitive regions), orthologous gene models might be truncated in one or more of the references. To our knowledge, a comprehensive evaluation of methods that could be applied to account for inter-specific gene length differences has not been performed yet.

A plethora of statistical approaches have been developed to determine whether differences in the number of reads are due to technical variation or due to real biological differences in gene expression. Detailed evaluation and comparison of these methods concluded that the most accurate statistical validation of differential gene expression is reached when statistical models are used that directly take the number of aligned reads into account (Bullard et al., 2010; Dillies et al., 2012; Rapaport et al., 2013; Sonesson and Delorenzi, 2013). These methods include standard and generalized Poisson and negative binomial distributions to model count-based expression data (Chu et al., 2015; Dillies et al., 2012) as implemented in DESeq (Anders and Huber, 2010), DESeq2 (Love et al., 2014b), edgeR (Robinson et al., 2010) or deGPS (Chu et al., 2015). Also the differential expression analysis based on moderated t-statistics as implemented in the limma package (Ritchie et al., 2015; Smyth,

2004) using log-transformed count per million values originating from normalization with voom (Law et al., 2014) (referred to as voom-limma below) has been shown to perform extremely well (Rapaport et al., 2013). While all of these methods account for most technical biases and highly reduce the false positive rate, none of these methods is specifically designed to account for gene length differences as they occur in inter-specific expression studies. One potential solution could be the application of the normalization method reads per kilobase per million mapped reads (RPKM) as it accounts for length differences in gene models (Mortazavi et al., 2008). However, it has been shown that even after correcting for length differences, a longer transcript is more likely to appear as differentially expressed if RPKM values are used to assess the statistical significance (Bullard et al., 2010; Oshlack and Wakefield, 2009; Rapaport et al., 2013; Robinson and Oshlack, 2010; Wagner et al., 2012). RPKM normalization is still widely used to compare gene expression levels of different genes within a species, but to our knowledge it has not been tested if this method efficiently normalizes length differences when comparing gene expression in different species.

Here we show that the published genomes of three closely related *Drosophila* species, *D. melanogaster*, *D. simulans* and *D. mauritiana* have qualitative limitations as references for comparative gene expression studies. This is mainly due to the fact that many genes cannot be properly compared because orthologous genes are missing in the annotation of at least one of the genomes. Even after a direct re-annotation of the three genomes using the same annotation pipeline many orthologous gene models exhibit significant length differences. Taking advantage of these inter-specific gene length differences in the published and the directly re-annotated references, we benchmarked four statistical frameworks (DESeq2 without length correction, DESeq2 with length correction, RPKM-limma and RPKM-voom-limma) for their ability to reduce the number of potentially false positives. We demonstrate that none of these methods sufficiently accounts for the observed differences in gene length. Therefore, we propose that the length normalization should be performed prior to read mapping during the generation of the mapping references. We report a straightforward re-annotation method that relies on a reciprocal re-annotation of orthologous gene models in two or more species. This approach allows the comparison of nearly all genes that have been annotated in *D. melanogaster* in all three species. Additionally, we find that the use of the new gene sets as mapping references results in a more robust estimation of transcript abundance and a more reliable comparison of gene expression levels between species. We propose that the generation and annotation of new genome

resources or the re-annotation of existing genomes will be powerful tools to establish gene expression profiling in many emerging model systems.

4.2.3 Results and Discussion

4.2.3.1 Analysis of published genome annotations reveals a reduced number of comparable gene models for differential gene expression studies between species

We first assessed the completeness and comparability of the published gene sets for the three closely related species *D. melanogaster*, *D. simulans* and *D. mauritiana*. At the time of our analysis, the annotation of the *D. melanogaster* genome (r5.55) - one of the best curated metazoan genomes available at FlyBase (Adams et al., 2000; Myers et al., 2000; St. Pierre et al., 2014) - included 13,676 unique protein coding genes. The most recent annotations for *D. simulans* (Hu et al., 2013) and *D. mauritiana* (Nolte et al., 2013) were generated using the *D. melanogaster* gene set as reference (for the *D. simulans* project the authors used the *D. melanogaster* annotation r5.33, and for the *D. mauritiana* project r5.32 was used). Both gene sets contain a large fraction of the 13,676 *D. melanogaster* genes (86.55% in *D. simulans* and 87.78% in *D. mauritiana*, Table 4.2.1). However, orthologs of almost 2,000 *D. melanogaster* genes are not included in each of the final gene sets either because the authors applied various filtering steps to exclude incomplete orthologous sequence with respect to the *D. melanogaster* gene (see the filtering criteria in the Methods of (Hu et al., 2013; Nolte et al., 2013)) or because the genes are not present in one of the species. Since these filtering steps are influenced by the quality of each of the assembled genome and the scientific question of each research group, the missing genes in both annotations are not the same. Only 9,994 genes (73.08%) can be identified unequivocally as orthologs in all three annotations (see Materials and Methods). Among the genes missing in at least one annotation, we found some important and well-studied developmental genes including the Hox genes *abdominal B* (*abd-B*), *Ultrabithorax* (*Ubx*) or *Antennapedia* (*Antp*), the head and brain patterning gene *orthodenticle* (*otd*) and the segment polarity gene *hedgehog* (*hh*) (data not shown).

Table 4.2.1. Number of genes obtained by each annotation method.

Method	<i>D. melanogaster</i>	<i>D. simulans</i>	<i>D. mauritiana</i>	Comparable
Published annotation after filtering	13,676	11,837 (86.55%)	12,005 (87.78%)	9,994 (73.08%) 8,810 (64.42%)
Direct re-annotation after filtering	13,676	13,436 (98.24%)	13,401 (97.99%)	13,328 (97.45%) 12,334 (90.19%)
Reciprocal re-annotation after filtering	13,457 (98.40%)	13,373 (97.78%)	13,346 (97.59%)	13,311 (97.33%) 13,239 (96.80%)

The last column contains the number of genes for which 1:1 orthologs were identified in the three species. “after filtering” indicates the remaining common genes after filtering out genes with length difference larger than 49 bp. Percentages in brackets are always given in relation to the total number of gene models in *D. melanogaster* (r5.55; 13,676 gene models).

Next we assessed the comparability of the three reference genome annotations in terms of gene length, since length differences larger than the length of the RNA-seq reads are likely to introduce a bias during mapping and the subsequent differential expression analysis. If we consider 50 bp single-end reads, which have been shown to be long enough to produce accurate results when measuring differential gene expression (Chhangawala et al., 2015; González and Joly, 2013; Li and Dewey, 2011), genes that have a length difference larger than 49 bp among the annotations of the three *Drosophila* species are likely to bias a subsequent differential gene expression analysis. A pair-wise comparison of annotated gene length for the 9,994 genes present in all three *Drosophila* species (Figure 4.2.1A) shows that in the published annotations, the gene length differences are larger than 49 bp in 7.6% (757) of the orthologous genes between *D. mauritiana* and *D. simulans*, 9.1% (912) between *D. melanogaster* and *D. simulans* and 7.1% (706) between *D. melanogaster* and *D. mauritiana* (Figure 4.2.1A, Supplementary Table 2). If these length differences are not accounted for, these genes could result in false positives in a differential gene expression analysis.

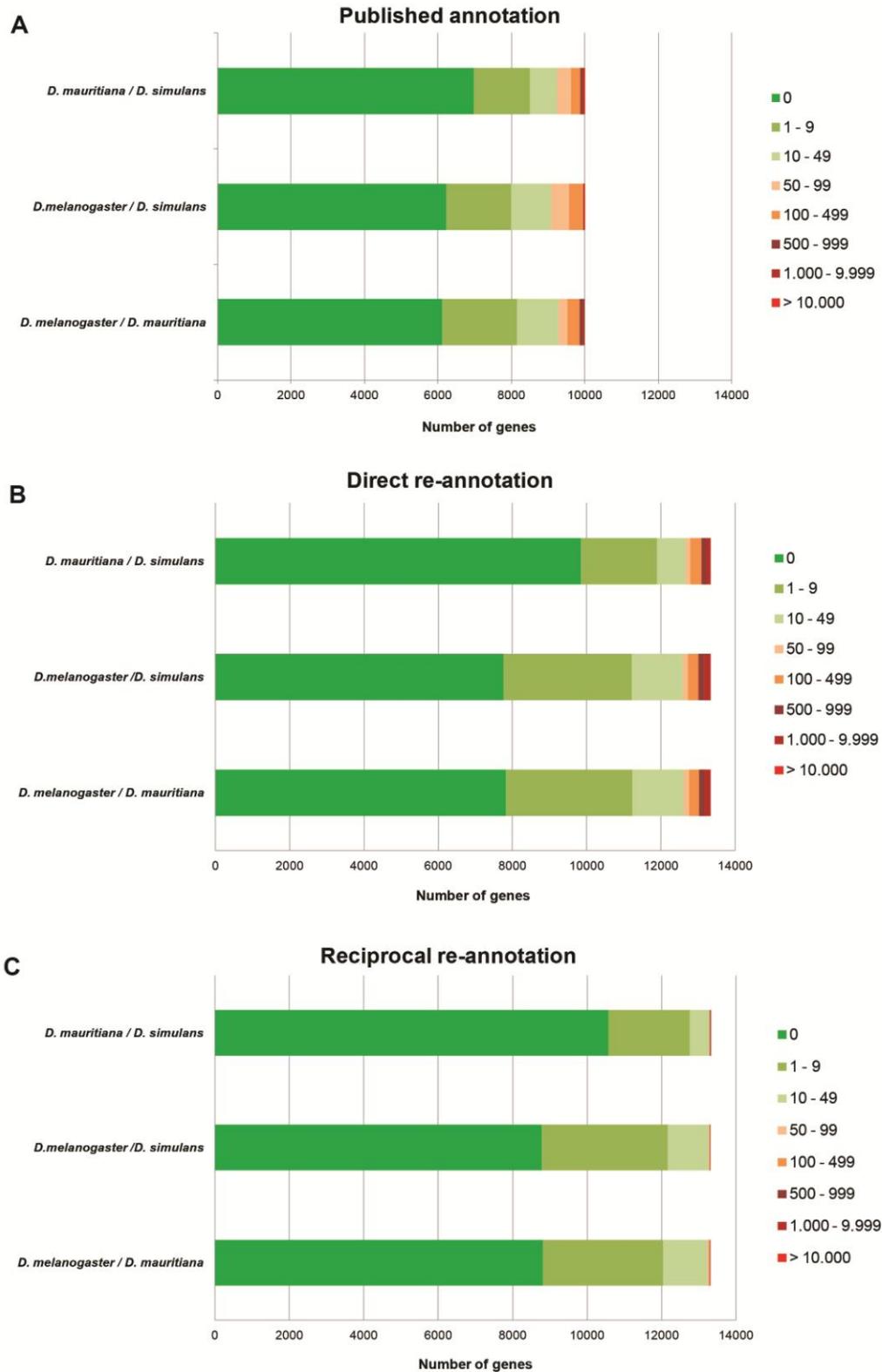


Figure 4.2.1. Pair-wise length difference between orthologous genes. Bars indicate the number of genes with that difference in length (calculated in number of nucleotides in the annotated transcripts) for each pair of species. Green shades indicate differences lower than 50 bp while orange to red indicate larger differences. The comparison is showed for **(A)** the published annotations, **(B)** the direct re-annotation of the published genomes and **(C)** the reciprocal re-annotation of the published genomes.

4.2.3.2 Direct re-annotation of published genomes

Next we asked whether a direct re-annotation of the *D. simulans* and *D. mauritiana* genomes individually using the same *D. melanogaster* gene set as reference and the same annotation pipeline allows the comparison of more genes in an inter-specific differential gene expression study.

We used the 13,676 protein sequences of *D. melanogaster* (r.5.55) as reference to re-annotate the published genomes of *D. simulans* (Hu et al., 2013) and *D. mauritiana* (Nolte et al., 2013) using the program Exonerate (Slater and Birney, 2005). Without applying any filtering, we find orthologs of 13,328 *D. melanogaster* genes that are comparable among the two species (Table 4.2.1). Next, we determined the length of directly re-annotated genes that are found in all three species. This comparison shows an increase in number of genes with the same length between the three species after direct re-annotation (Figure 4.2.1B; Supplementary Table 2). However, a high number of orthologous genes have a length difference of more than 49 bp (Figure 4.2.1B; Supplementary Table 2): 706 (5.3%) between *D. melanogaster* and *D. mauritiana*, 740 (5.6%) between *D. melanogaster* and *D. simulans* and 658 (4.9%) between *D. mauritiana* and *D. simulans*. These observed length differences could be due to real natural variation in coding sequence length between species or they could be technical artifacts, for example truncated gene sequences arising from sequencing or genome assembly errors.

In summary, although annotated genomes are available for the three closely related *Drosophila* species *D. melanogaster*, *D. simulans* and *D. mauritiana*, their use as mapping references for inter-species differential gene expression analyses is limited due to missing orthologs and a potential bias because of different annotated gene lengths. The use of the same reference gene set, annotation pipeline and the lack of filtering incomplete gene sequences results in an increase in the number of comparable genes in these three closely related species.

4.2.3.3 Length difference in reference genes introduces biases in differential expression studies

Since we find a high number of gene models with length differences > 49 bp in the published annotations and after the direct re-annotation (Figure 4.2.1A and B; Supplementary Table 2), the three *Drosophila* genomes are excellent models to test whether length differences larger than the read length do indeed influence the statistical analysis of differential gene expression. We used the published *D. melanogaster* annotation and the newly generated direct re-annotations of *D. simulans* and *D. mauritiana* as mapping

references for pair-wise comparisons of gene expression between *D. melanogaster* and *D. mauritiana* and *D. simulans* and *D. mauritiana* using 50 bp single-end Illumina RNA-seq reads generated for these three species (see Materials and Methods). The mapping was always species-specific: RNA-seq reads generated from one species were mapped only to the gene set of that species.

Table 4.2.2. Differentially expressed genes and correlation between calculated log₂-fold changes and length difference between orthologous genes.

Method	Annotation	Species	# Common genes	# Differentially expressed genes (% of common genes)*	Spearman's p value ⁺	FDR 0.05	Spearman's rho ⁺
DESeq2	Published	<i>D. mau</i> vs <i>D. mel</i>	11,503	2,438 (21.2)	6.52e-33	***	-0.36
		<i>D. mau</i> vs <i>D. sim</i>	10,023	2,974(29.7)	2.15e-20	***	-0.33
	Direct	<i>D. mau</i> vs <i>D. mel</i>	13,401	2,665 (19.9)	3.35e-20	***	-0.33
		<i>D. mau</i> vs <i>D. sim</i>	13,328	3,710 (27.8)	5.90e-58	***	-0.57
	Reciprocal	<i>D. mau</i> vs <i>D. mel</i>	13,331	2,501 (18.8)	5.12e-02	n.s.	-0.23
		<i>D. mau</i> vs <i>D. sim</i>	13,320	3,508 (26.3)	1.48e-01	n.s.	-0.29
DESeq2 + length matrix	Published	<i>D. mau</i> vs <i>D. mel</i>	11,503	1,192 (10.4)	2.24e-05	***	-0.13
		<i>D. mau</i> vs <i>D. sim</i>	10,023	1,545 (15.4)	3.20e-02	*	-0.08
	Direct	<i>D. mau</i> vs <i>D. mel</i>	13,401	1,259 (9.4)	1.07e-02	*	-0.09
		<i>D. mau</i> vs <i>D. sim</i>	13,328	1,957 (14.7)	5.24e-04	**	-0.13
	Reciprocal	<i>D. mau</i> vs <i>D. mel</i>	13,331	1,215 (9.1)	7.03e-01	n.s.	-4.6e-02
		<i>D. mau</i> vs <i>D. sim</i>	13,320	1,910 (14.3)	7.34e-01	n.s.	-0.07
RPKM + limma	Published	<i>D. mau</i> vs <i>D. mel</i>	11,503	1,904 (16.6)	4.42e-04	***	-0.11
		<i>D. mau</i> vs <i>D. sim</i>	10,023	2,427 (24.2)	1.06e-03	**	-0.12
	Direct	<i>D. mau</i> vs <i>D. mel</i>	13,401	1,890 (14.1)	5.68e-03	*	-0.10
		<i>D. mau</i> vs <i>D. sim</i>	13,328	2,795 (21.0)	4.49e-04	***	-0.14
	Reciprocal	<i>D. mau</i> vs <i>D. mel</i>	13,331	1,830 (13.7)	5.92e-01	n.s.	-6.4e-02
		<i>D. mau</i> vs <i>D. sim</i>	13,320	2,738 (20.6)	2.83e-01	n.s.	-0.22
RPKM + voom + limma	Published	<i>D. mau</i> vs <i>D. mel</i>	11,503	1,853(16.1)	9.39e-04	***	-0.10
		<i>D. mau</i> vs <i>D. sim</i>	10,023	2,204(22.0)	4.63e-02	*	-0.07
	Direct	<i>D. mau</i> vs <i>D. mel</i>	13,401	1,899(14.2)	1.01e-02	*	-0.10
		<i>D. mau</i> vs <i>D. sim</i>	13,328	2,607(19.6)	5.92e-03	*	-0.11
	Reciprocal	<i>D. mau</i> vs <i>D. mel</i>	13,331	1,819(13.6)	5.79e-01	n.s.	-0.07
		<i>D. mau</i> vs <i>D. sim</i>	13,320	2,519(18.9)	4.06e-01	n.s.	-0.17

Results are shown for the four applied methods, the three studied annotation strategies and the two described pair-wise species comparisons.

* FDR 0.05

⁺Spearman's rank correlation is measured between log₂FC and length difference of genes showing more than 49bp length difference: Published annotation: *D. mau* vs. *D. mel*: 1,038 genes / *D. mau* vs. *D. sim*: 764 genes; Direct annotation: *D. mau* vs. *D. mel*: 716 genes / *D. mau* vs. *D. sim*: 658 genes; Reciprocal annotation: *D. mau* vs. *D. mel*: 71 genes / *D. mau* vs. *D. sim*: 26 genes

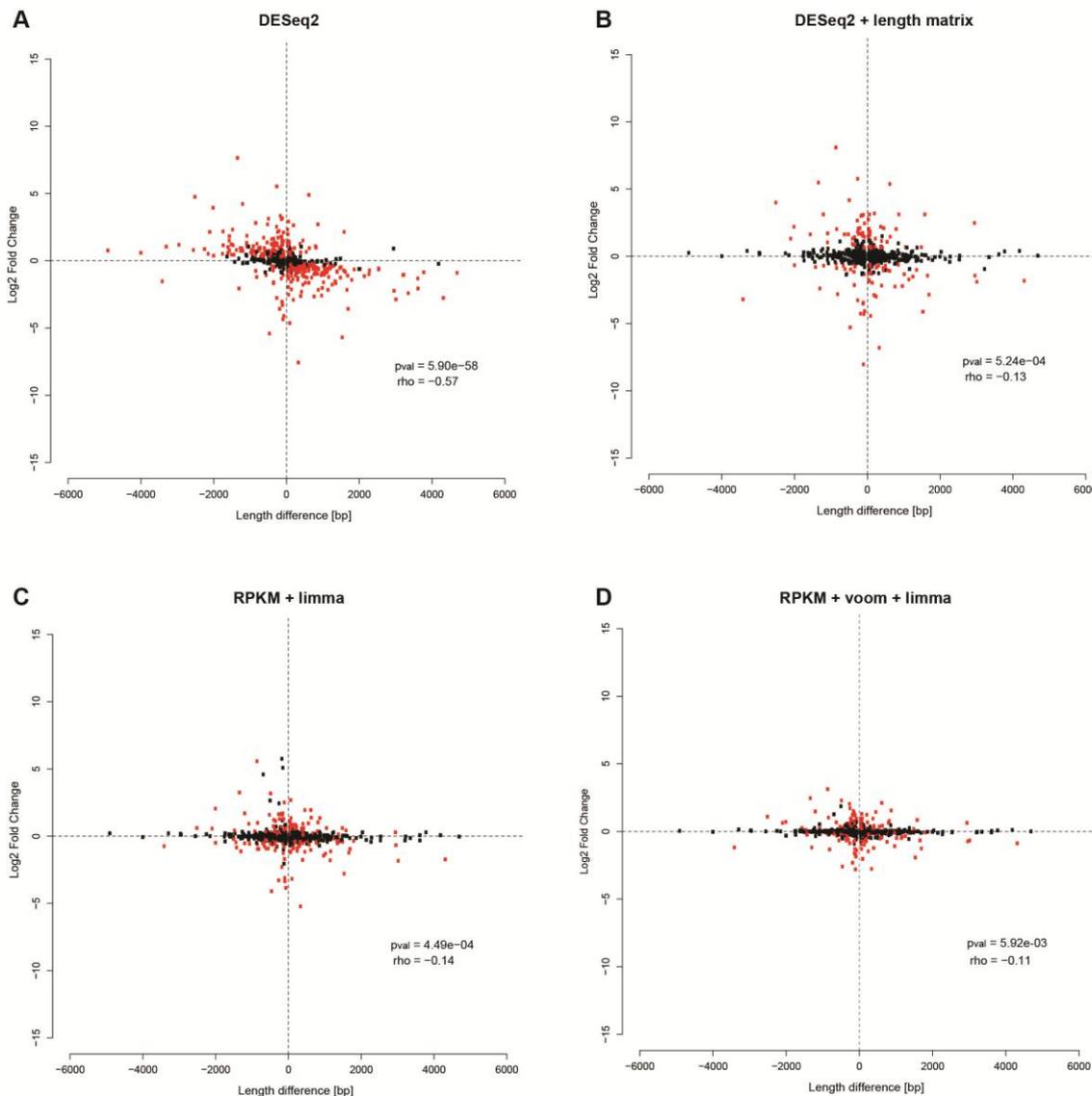


Figure 4.2.2. Length differences between orthologous genes introduce gene expression biases. Relation between length differences and the log₂-fold change in the RNA-seq experiment between *D. mauritiana* and *D. simulans* using the direct re-annotation of these species as mapping references. Dots represent genes with length difference > 49 bp in these annotations (658 genes). Genes significantly differentially expressed in the presented analysis ($p\text{-adj} < 0.05$) are shown in red. A negative log₂-fold change indicates higher expression in *D. mauritiana*. A positive length difference indicates that the ortholog of *D. mauritiana* is longer. The p-value and rho of the Spearman's rank correlation are indicated on the lower right side of the plots. **(A)** Results of DESeq2 without length correction. **(B)** Results of DESeq2 applying length normalization factor matrix. **(C)** Results of applying RPKM normalization and limma to call differentially expressed genes. **(D)** Results of applying voom normalization followed by a length normalization matrix and limma to call differentially expressed genes.

Using this experimental setup, we compared four different statistical frameworks, namely DESeq2, DESeq2 with gene length correction (Love et al., 2014b), limma with length correction based on RPKM (Mortazavi et al., 2008; Ritchie et al., 2015; Russo and Angelini, 2014; Smyth, 2004) and voom-limma (Law et al., 2014) including RPKM length correction.

For each method, we first report the number of differentially expressed genes for each of the two pair-wise species comparisons. Next we evaluate the impact of the length differences between gene models on the fold-change in gene expression between species. And eventually, we compare the results of each of the four methods to an independent qPCR experiment for a subset of genes.

DESeq2 without length correction

First we performed the statistical analysis for differential gene expression with DESeq2 (Love et al., 2014b) using directly the read counts for each gene model. For both pair-wise comparisons using the published and the direct re-annotation as reference, we found that between 19.9% and 29.7% of all comparable genes are significantly differentially expressed (Table 4.2.2).

Additionally, we found a very strong correlation between inter-specific length differences of the gene models (considering only those gene models with differences > 49 bp) and the log₂-fold change in gene expression (Figure 4.2.2A, Supplementary Figure 4; Table 4.2.2). The negative correlation means that genes that are longer in one species appear to be more up-regulated and vice versa. The correlation can be explained by the mapping procedure: in orthologous genes with different length, more reads align to the ortholog that has the longer gene model (Figure 4.2.3, upper panel). This results in an artificially higher expression for this specific gene in the species with the longer gene model. From this correlation we also see that most of those genes with length differences and a high log₂-fold change are also significantly differentially expressed (Figure 4.2.2A, Supplementary Figure 4, $p < 0.05$, red dots), showing that this method introduces a large number of false positives.

In order to specifically test whether differences in the length of gene models indeed influence the differential expression analysis we chose seven genes that were shorter in the *D. mauritiana* published annotation compared to the published *D. melanogaster* annotation. When we analysed the differential expression using DESeq2 without any length correction, the expression of all seven genes were significantly different (Table 4.2.3). The log₂-fold change value indicated that *D. melanogaster* had a significantly higher expression than *D. mauritiana* (Table 4.2.3). To validate the results obtained by the RNA-seq experiment, we measured the relative expression of the seven genes in *D. melanogaster* and *D. mauritiana* using qPCR. This analysis showed that the seven genes that had length differences in the species-specific annotations were not significantly differentially expressed (Figure 4.2.4,

Table 4.2.3). As a control we chose another three genes that showed significant differential expression in the RNA-seq data but had the same length in both species in the two annotation methods (*pivi* and *alrm* are significantly higher in *D. mauritiana* and *Nplp1* is higher in *D. melanogaster*). We found that *pivi* and *alrm* showed a significantly higher expression in *D. mauritiana* when using this alternative quantification method, confirming the results obtained by RNA-seq (Figure 4.2.4, Table 4.2.3). Moreover, *Nplp1* had higher expression in *D. melanogaster* again consistent with our RNA-seq data, although this difference was not significant ($p=0.072$; Figure 4.2.4, Table 4.2.3).

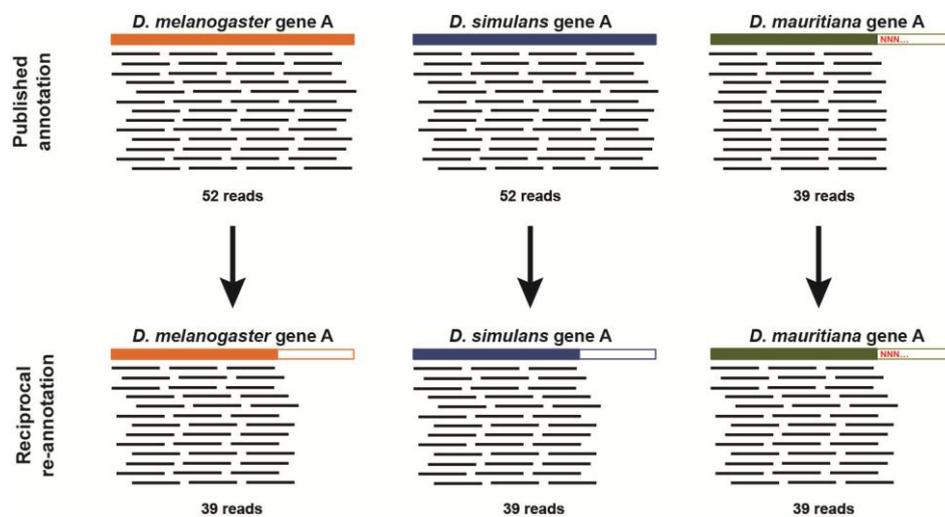


Figure 4.2.3. Schematic representation of length bias in inter-species differential expression analysis and our reciprocal re-annotation strategy to correct it. Length bias in the analysis of a non-differentially expressed gene. Colored rectangles represent the part of the transcript which is included as reference for the RNA-seq reads to map to, while unfilled rectangles are regions of the transcript which are omitted and to which RNA-seq reads cannot be mapped. Red “N”s represent sequencing errors that prevent the complete annotation of a transcript. Mapped reads are shown as thin black lines and the number below indicates the total of reads mapped. **(upper panel)** If one transcript is shorter in one of the references compared to its orthologs, for the same expression levels fewer reads will map to it. This can result in false positives in the analysis of differential expression. **(lower panel)** Our strategy to correct this bias is to shorten the orthologs in the other references to match the length of the shorter sequence.

Table 4.2.3. Analysis of differential expression. Expression comparison is for *D. mauritiana* vs. *D. melanogaster*, thus a positive log₂-fold change (log₂FC) indicates higher expression in *D. melanogaster* and vice versa. *: $p < 0.05$; **: $p < 0.005$; ***: $p < 0.0005$.

Gene	qPCR log ₂ FC	Published transcriptomes						Reciprocally re-annotated transcriptomes		
		Gene length (# nucl.)		DESeq2	DESeq2 + length matrix	RPKM + limma	RPKM + voom + limma	Gene length (# nucl.)		DESeq2
		<i>D. mel</i>	<i>D. mau</i>	log ₂ FC	log ₂ FC	log ₂ FC	log ₂ FC	<i>D. mel</i>	<i>D. mau</i>	log ₂ FC
lace	-0.19	1791	903	1.40***	0.38	0.41	0.15	902	902	0.03
CG3558	0.08	3147	1956	1.50***	0.67	0.75	0.26*	3150	3135	0.16
dac	-0.29	3243	1878	1.47***	0.57	0.65	0.18	1887	1878	0.46
RAF2	1.0e-03	3351	1854	1.77***	0.84	0.94	0.31	1959	1966	0.33
Cp110	-0.18	1998	1218	2.31***	1.38*	1.4**	0.55**	2000	1998	0.11
CBP	-0.21	1653	894	1.42***	0.35	0.54	0.14	1656	1653	-0.24
CG6766	-0.41	1575	852	1.81***	0.79	0.88	0.25*	855	855	0.31
piwi	-2.60**	2529	2526	-2.48***	-2.54***	-1.99**	-1.08**	2532	2529	-2.48***
alrm	-2.37***	1413	1413	-6.54***	-6.67***	-4.93***	-2.68***	1416	1416	-6.49***
Nplp1	1.04	1461	1461	3.85***	3.63***	3.06***	1.50***	1464	1464	3.80***

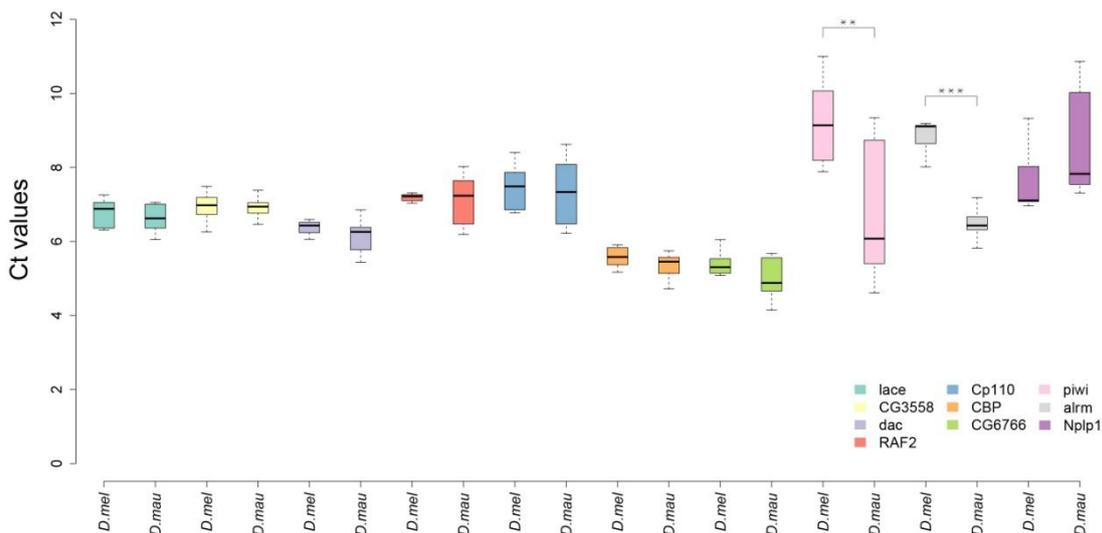


Figure 4.2.4. qPCR results. Boxplot of normalized Ct values (reference gene: *actin 79B*) For each studied gene (one color) boxplot is showed for Ct values in *D. melanogaster* OreR (“D. mel”) and *D. mauritiana* TAM16 (“D. mau”). (Significance calculated by t-test (for genes with homogeneous distribution of variances) or t-Welch-test (for genes with not homogeneous distribution of variances); **: $p < 0.05$, ***: $p < 0.005$; ****: $p < 0.0005$).

In summary, these results suggest a high level of potentially false positive candidates when methods based on direct read counts without the application of length correction are used with mapping references that exhibit differences in the length of orthologous genes.

DESeq2 with length correction

Next we benchmarked the use of DESeq2 (Love et al., 2014b) including a normalization factor matrix incorporating gene length to account for the length differences between orthologous genes. Using this approach for pair-wise gene expression analyses, we found that only 9.4% to 15.4% of the comparable genes were significantly differentially expressed. Even though the gene length was accounted for during the DESeq2 analysis of differential gene expression, we still find a correlation between inter-specific gene length differences and log₂-fold changes (Figure 4.2.2B, Supplementary Figure 5, Table 4.2.2). However, the significance of this correlation is greatly reduced in comparison to the DESeq2 analysis without length correction (Table 4.2.2), suggesting that the length correction incorporated in DESeq2 helps to reduce the number of false positive candidate genes. This finding is further supported by the comparison of the RNA-seq results to the qPCR data. After length correction only one (*Cp110*) of the seven genes that are longer in *D. melanogaster* show significant differential expression (Figure 4.2.4, Table 4.2.3).

limma with RPKM length correction

RPKM values are commonly calculated for RNA-seq datasets to account for variation in library sizes and to correct for length differences between different genes within the same species (Mortazavi et al., 2008). The moderated t-statistics incorporated in the limma R package (Ritchie et al., 2015; Smyth, 2004) can subsequently be used to assess differential gene expression. It has not been tested if this approach also corrects properly for differences in the length of the same gene being compared between two species. Using this method, we found between 14.1% and to 24.1% of the comparable genes to be significantly differentially expressed (Table 4.2.2). Our correlation analysis shows that the correction of a length bias using RPKM values still results in a clearly significant correlation between the gene length difference and the observed log₂-fold change (Figure 4.2.2C, Supplementary Figure 6, Table 4.2.2). However, compared to the DESeq2 analysis without length correction, the significance values are highly reduced (Table 4.2.2), showing that the number of false positives is lower. Accordingly, six of the seven genes that we benchmarked with qPCR show no significant differential gene expression although they show clear length differences between *D. melanogaster* and *D. mauritiana* (Table 4.2.3). Again

Cp110 is the only gene that appears as significantly differentially expressed also after correcting for length differences.

voom-limma with RPKM length correction

It has recently been shown that differential gene expression analysis with limma (Ritchie et al., 2015; Smyth, 2004) using normalized read counts from voom (Law et al., 2014) perform very well for RNA-seq datasets (Rapaport et al., 2013). Although this method is designed to work with direct read counts, in this case we tested it with an additional transcript length correction. Between 15% and 23.5% of the comparable genes are significantly differentially expressed (Table 4.2.2). After length correction (RPKM) and normalization with voom, we found a significant correlation between gene length differences and log₂-fold changes when the published annotations and the directly re-annotated reference gene sets were used. However, this was slightly reduced compared to the RPKM-limma analysis, especially for the *D. simulans* and *D. mauritiana* comparison. (Figure 4.2.2D, Supplementary Figure 7, Table 4.2.2). For the seven qPCR benchmarked genes that have clear length differences between *D. melanogaster* and *D. mauritiana* the use of the voom-limma method results in three significantly differentially expressed genes (Table 4.2.3), suggesting a higher false positive rate.

Length correction during the statistical analysis might be insufficient

The comprehensive comparison of four methods for differential gene expression analysis shows that the incorporation of a length correction drastically reduces the number of false positive candidate genes. Although the correlation between gene length differences and the observed log₂-fold changes (Figure 4.2.2, Supplementary Figure 4-7, Table 4.2.2) is reduced in the three methods that account for gene length differences (length matrix in DESeq2, RPKM-limma and RPKM-voom-limma), none of them sufficiently corrects the length bias present in the two gene sets used as mapping references. This is also supported by the qPCR validation of seven genes that exhibit clear length differences between the published *D. melanogaster* and *D. mauritiana* annotations (Figure 4.2.4, Table 4.2.3). In all three methods at least one gene was still artificially significantly differentially expressed. This is most pronounced for the voom-limma method where three of the seven genes are significantly differentially expressed. Of the seven genes we analyzed using qPCR, *Cp110* was in all cases identified as a false positive candidate. In order to further characterize this gene, we visually inspected the distribution of mapped reads. Interestingly, the 3'-region is missing in the *D. mauritiana* ortholog of *Cp110* (Supplementary Figure 8A) and we found clearly more *D.*

melanogaster reads that map to this 3'-part of the transcript than to the 5'-region (Supplementary Figure 8C). The number of *D. melanogaster* and *D. mauritiana* reads mapped to the 5'-part of the transcript is comparable (compare Supplementary Figure 8B to C). Hence, a very likely explanation for the inefficient length correction of the three applied methods could be an unequal distribution of the mapped reads along the transcripts.

Besides an insufficient length correction, the DESeq2 method including a length matrix results in the lowest number of significantly differentially expressed genes, suggesting that the length correction applied here might be extremely conservative and could lead to a high rate of false negatives. Interestingly, in many pair-wise comparisons we found more significantly differentially expressed genes with a higher expression in *D. mauritiana* compared to *D. melanogaster* and *D. simulans* (not shown), although *D. mauritiana* gene models are generally shorter than those of the other two species (Figure 4.2.2, Supplementary Figure 4-7). This finding suggests that the length correction applied here might reduce the power to detect differential expression for the already short *D. mauritiana* genes.

In summary, all three methods that include a length correction decrease the chance of identifying false positives. The RPKM-voom-limma and RPKM-limma methods seem to give the best ratio of false positives and false negatives, while DESeq2 including a length matrix is very conservative. However, none of the length correction methods tested does efficiently account for all differences in gene length observed in the reference annotation of the three studied *Drosophila* species. The length bias is most obvious when the distribution of reads is not uniform across the transcript body (e.g. *Cp110*). Therefore, all genes that exhibit length differences larger than the read length should be excluded from any of the reference gene sets (see Table 4.2.1; number of comparable genes after filtering).

4.2.3.4 Reciprocal re-annotation reduces the number of false positive candidates

Overview of the reciprocal re-annotation pipeline

To overcome problems due to length differences between orthologous genes and simultaneously maximize the number of comparable genes, we developed a pipeline to reciprocally re-annotate the reference genomes of the three species (Figure 4.2.5, Materials and Methods). Instead of directly annotating the *D. simulans* and *D. mauritiana* genomes individually using the *D. melanogaster* reference gene set, we first annotated the genome of *D. simulans* based on the protein set from *D. melanogaster*. Subsequently, we used these newly

annotated *D. simulans* gene models to annotate the genome of *D. mauritiana*. This gene set was then used as a reference to re-annotate again the previously generated *D. simulans* gene set. And finally, we used these *D. simulans* gene models that already contain consensus features of *D. simulans* and *D. mauritiana* to re-annotate the *D. melanogaster* gene set (Figure 4.2.5). Therefore, we obtained the longest sequence present in all three species and then, if necessary, reduce its length in the other references accordingly. Thus, we expect to equalize the length of all the genes for the three references (Figure 4.2.3, lower panel). It is important to note here that it does not matter in which order the reciprocal re-annotation is done. As long as the first reference is the one of *D. melanogaster* (i.e. the best curated annotation), we obtained the same results when we first annotate *D. simulans* or *D. mauritiana* (not shown).

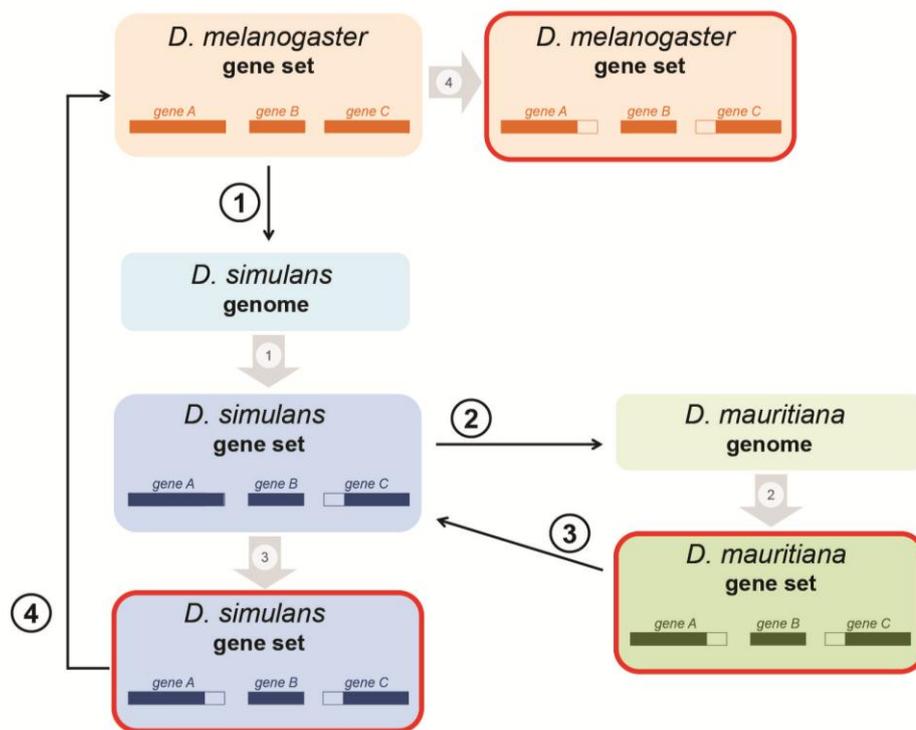


Figure 4.2.5. Pipeline of reciprocal transcriptome re-annotation method. Black numbers in white circles represent genome annotation steps using the “est2genome” command of Exonerate (Slater and Birney, 2005). Grey numbers in grey circles represent conversion of the resulting GFF file into a new transcript set. Filled horizontal bars represent the annotated set of transcripts; non-filled horizontal bars at the start/end of the transcripts represent parts of the transcript that cannot be correctly annotated in one reference and are therefore eliminated from the transcript set. The boxes with red frame indicate the transcript sets that will be used as reference for RNA-seq read mapping (after confirmation by reciprocal blast). **Step 1:** the transcript set of the best annotated genomes (*D. melanogaster* in our study) is used to annotate one of the other genomes (*D. simulans* in our study) and generate a new transcript set for this species. Due to sequencing errors, some transcripts will be shorter. **Step 2:** the new transcript set from *D. simulans* is used to annotate the last genome (*D. mauritiana* in our study). The gene set generated contains shorter transcripts due to sequencing errors in *D. mauritiana* but also in *D. simulans*. **Step 3:** the transcript set from *D. mauritiana* is used to re-annotate the previously generated set from *D. simulans* to integrate the

information from the *D. mauritiana* assembly. **Step 4:** the second transcript set from *D. simulans* is used to annotate the *D. melanogaster* set in order to integrate the information from *D. simulans* and *D. mauritiana*.

Reciprocal re-annotation efficiently reduces gene length differences between species

With the reciprocal re-annotation of the published genomes we obtained 97.33% of the 13,676 *D. melanogaster* gene models in each of the three species (Table 4.2.1). In accordance with our expectations, only a small fraction of those genes found in all three species have a length difference of more than 49 bp (Figure 4.2.1C; Supplementary Table 2): 71 genes (0.53%) genes between *D. melanogaster* and *D. mauritiana*, 41 genes (0.3%) between *D. melanogaster* and *D. simulans* and only 26 genes (0.19%) between *D. mauritiana* and *D. simulans*. Hence, the reciprocal re-annotation of the published genomes allows the analysis of the highest number of comparable genes with less than 50 bp length differences in a differential gene expression study (Table 4.2.1; 13,239 (96.80%) of the 13,676 *D. melanogaster* gene models).

Evaluation of the reciprocal re-annotation in RNA-seq experiments

To quantitatively test whether the number of false positives due to gene length differences is indeed reduced after reciprocal re-annotation, we applied a pair-wise analysis of differential gene expression between *D. melanogaster* and *D. mauritiana* and *D. simulans* and *D. mauritiana* (see Materials and Methods). We mapped the RNA-seq reads to the new references and performed the statistical analysis using the four methods evaluated above: DESeq2, DESeq2 with length correction (Love et al., 2014b), RPKM-limma (Mortazavi et al., 2008; Ritchie et al., 2015; Smyth, 2004), RPKM-voom-limma (Law et al., 2014).

As for the published and directly re-annotated references, the statistical analysis with DESeq2 resulted in the highest number of significantly differentially expressed genes (18.8% and 26.3% of the comparable genes; Table 4.2.2). This number clearly dropped to 9.1% and 14.3% after including a length correction during the DESeq2 analysis. Similarly, the number of differentially expressed genes is greatly reduced if RPKM-limma and RPKM-voom-limma are being used (Table 4.2.2). However, only 71 (*D. melanogaster* vs. *D. mauritiana*) and 26 (*D. simulans* vs. *D. mauritiana*) pair-wise comparable genes exhibit length differences greater than 49 bp after reciprocal re-annotation. One would expect that only those genes should be affected by any of the three length correction methods.

Therefore, we propose that the combination of a reciprocal re-annotation in combination with a read-count-based DESeq2 analysis of differential gene expression is likely to provide

the most comprehensive and reliable estimation of inter-specific gene expression differences. This is further supported by the lack of a significant correlation between log₂-fold changes and gene length differences if the DESeq2 is used in combination with the reciprocal re-annotation as mapping reference (Supplementary Figure 4-7; Table 4.2.2). Although the correlation is not significant, we still find that most significantly differentially expressed genes with length differences larger than 49 bp have higher expression in the species with the longer transcript (Supplementary Figure 4-7). Therefore, we propose that those genes should be filtered out from further differential gene expression analysis. Additionally, the seven genes that were validated using qPCR did not show a significant differential expression after their length was equalized (Figure 4.2.4, Table 4.2.3), suggesting that the length correction during the annotation of genomes can facilitate the reduction in false positive candidate genes in RNA-seq analyses.

Assessment of power to detect differential gene expression

It is important to note that the gene models generated by our reciprocal re-annotation pipeline do not necessarily represent the complete gene and thus the most comprehensive annotation for each species. This is due to the fact that potential full gene models in one species might have been adjusted to the shortest orthologous gene model. Therefore, in each round of annotation some gene models are truncated to fit the length of its orthologs (see Figure 4.2.3 and 4.2.2). If the gene models would be extremely shortened, this could of course lead to a loss of statistical power for the differential gene expression analysis. In order to estimate how much sequence information we really lose, we compared the length of the *D. melanogaster* gene models before and after the reciprocal re-annotation. This comparison shows that 12,642 (92.44%) of the 13,676 gene models still contain 90% to 100% of their original sequence length after the reciprocal re-annotation (Supplementary Figure 9).

Next we assessed the potential loss of power by comparing the number of mapped reads between the published annotations (13,676 genes in *D. melanogaster*, 12,005 genes in *D. mauritiana* and 11,837 genes in *D. simulans*; Table 4.2.1) and the gene sets generated from our reciprocal re-annotation of the published genomes (13,457 genes in *D. melanogaster*, 13,373 genes in *D. simulans* and 13,346 genes in *D. mauritiana*, Table 4.2.1). Overall, the proportion of successfully mapped reads for all reference gene sets was between 40% and 67% (Table 4.2.4). A large portion of this relatively low mapping rate can be explained by the fact that we excluded UTR sequences from all reference gene sets, what accounts for

about 27.4% of all mapped reads (see Materials and Methods; Supplementary Table 3). Additionally, we only used the longest isoform of *D. melanogaster* for all annotations in the other two species (see Materials and Methods). Therefore, some differentially spliced exons might not be represented in the newly generated gene sets. However, the use of the sum of all exons only increases the mapping success by 0.4% if UTRs are excluded and 1.6% if the UTRs are included (Materials and Methods; Supplementary Table 3). If the comparison of the expression of different isoforms across species is of interest one could perform the quantification on the level individual transcripts (Soneson et al., 2015) or even exons. This approach requires of course a proper annotation of the different isoforms in all reference genomes and a dedicated mapping pipeline. For our analysis, we found for all replicates more than 17 million mapped reads after reciprocal re-annotation (Table 4.2.4) what has been shown to provide enough statistical power for differential gene expression analyses (Malone and Oliver, 2011).

Table 4.2.4. List of RNA-seq samples and the percentage and number of mapped reads to different reference transcriptomes.

Sample	Original read type*	Published transcriptomes		Reciprocally re-annotated transcriptomes	
		Percentage	Total number of mapped reads	Percentage	Total number of mapped reads
<i>D. melanogaster</i> replicate A	SE 50bp	58.86%	28,486,024	57.33%	27,744,730
<i>D. melanogaster</i> replicate B	SE 50bp	44.23%	17,675,472	43.19%	17,260,775
<i>D. melanogaster</i> replicate C	SE 50bp	65.51%	25,316,846	63.91%	24,699,746
<i>D. mauritiana</i> replicate A	SE 50bp	40.70%	16,575,011	43.31%	17,639,874
<i>D. mauritiana</i> replicate B	SE 50bp	56.17%	31,884,442	60.07%	34,100,435
<i>D. mauritiana</i> replicate C	SE 50bp	53.01%	23,653,723	56.98%	25,425,486
<i>D. mauritiana</i> replicate D	PE 100bp	56.06%	111,643,922	61.07%	121,610,905
<i>D. mauritiana</i> replicate E	PE 100bp	54.28%	130,638,956	59.51%	143,226,939
<i>D. mauritiana</i> replicate F	PE 100bp	60.90%	144,541,354	66.21%	157,165,639
<i>D. simulans</i> replicate A	PE 100bp	62.26%	118,272,529	66.71%	126,741,807
<i>D. simulans</i> replicate B	PE 100bp	57.90%	138,364,665	62.56%	149,508,494
<i>D. simulans</i> replicate C	PE 100bp	56.32%	150,692,651	60.98%	163,168,587

* Reads originally 100bp paired-end (PE) were split in half to be 50bp each and treated as single-end (SE) reads.

We observed an increase in the mapping percentage of up to 5% in *D. simulans* and *D. mauritiana* when the reciprocally re-annotated gene sets are used as references (Table 4.2.4). This result shows that, although some gene models were now shorter, many genes that had been filtered out in the published genome annotations are actually expressed in these species. The use of the re-annotated gene set only slightly decreases the mapping success by 1% to 1.6% in *D. melanogaster* (Table 4.2.4), which is likely to be due to the artificial shortening of *D. melanogaster* gene models.

In summary, we show that the artificial shortening of transcripts after reciprocal re-annotation does not have a major impact on the power to detect differential gene expression.

Practical considerations

We demonstrate that the use of all annotated exons instead of the longest isoform of each gene model does not significantly increase the power to detect differential gene expression. In contrast, the inclusion of UTR regions for the reciprocal re-annotation will clearly increase the number of mapped reads (Supplementary Table 3) and hence the statistical power. However, the availability of UTR sequence information strongly depends on the quality of the annotation of the species to compare, since UTR and isoform predictions usually profit from the presence of RNA-seq data to be incorporated in the annotation pipeline. Additionally, the annotation of UTR regions might become more complicated if more distantly related species are studied, because UTR regions tend to evolve faster than coding region (Andolfatto, 2005).

Although we used very closely related species for our analysis here, we think that the presented reciprocal re-annotation is also applicable for genomes of more distantly related species. As a consequence of a higher sequence divergence between distantly related species, inter-specific gene length differences are likely to be more pronounced. If such genomes were used as mapping references, the direct use of length correction during the statistical analysis of differential gene expression might enhance the over-correction effect that we have demonstrated for the three presented methods. Additionally, if the gene lengths are very different between species, the length bias that has been reported for RPKM-based normalization approaches (Bullard et al., 2010; Oshlack and Wakefield, 2009; Rapaport et al., 2013; Robinson and Oshlack, 2010; Wagner et al., 2012) might be more pronounced. Therefore, we propose that the correction of the inter-specific length bias prior to read mapping using our reciprocal re-annotation pipeline should result in more

robust results. However, for more distantly related species, the reciprocal re-annotation is likely to result in more artificial shortening of the genes. Since this could reduce the power to detect differential gene expression, we propose to assess the length differences between species as we presented it here (Figure 4.2.1) prior to the sample preparation and sequencing and to adjust the coverage accordingly by generating more reads to increase sequencing depth.

In the presented case, at least one of the three *Drosophila* species represents a well-established model system with a high quality genome assembly and annotation. If this is available, the reciprocal re-annotation pipeline should of course be started with the highest quality annotation. If the annotation quality of all genomes similar the pipeline could be started with any of the studied species, since we showed that the direction of the reciprocal annotation does not influence the final result. However, if the quality of all annotations is comparably low, one should consider generating longer paired-end reads with higher coverage to first perform a de novo annotation with tools like AUGUSTUS (Stanke and Waack, 2003; Stanke et al., 2008) or BRAKER1 (Hoff et al., 2015) using those reads to train the respective algorithm. Subsequently, the generated RNA-seq reads can be used to assess differential gene expression using the reciprocally re-annotated references with length adjusted orthologous genes.

4.2.4 Conclusions

We have carried out a comprehensive comparison of the annotations of published genomes for the three closely related *Drosophila* species, *D. melanogaster*, *D. simulans* and *D. mauritiana*. This analysis reveals that different assembly strategies, annotation pipelines and filtering steps result in only a small fraction of genes that are comparable among all three species. A direct re-annotation of the *D. simulans* and *D. mauritiana* genomes using the same *D. melanogaster* reference gene set and the same annotation pipeline significantly improves the comparability of the gene sets. However, this direct re-annotation still results in length differences in many gene models between species. Based on these length differences between orthologous genes we tested four alternative methods to statistically assess differential gene expression using RNA-seq, namely DESeq2, DESeq2 with length correction, RPKM-limma and RPKM-voom-limma. We show that none of these methods sufficiently accounts for the inter-specific gene length differences what is evident by a high number of false positive differentially expressed genes. This finding is further supported by qPCR as an alternative transcript quantification method.

In order to further reduce the observed false positive rate, we argue that the length bias should be accounted for prior to the RNA-seq analysis during the generation of the mapping references. Therefore, we implemented a robust reciprocal re-annotation pipeline that allows the generation of highly comparable gene sets to serve as mapping references for inter-specific RNA-seq experiments. Applying RNA-seq and qPCR we confirm the successful reduction of false positive candidate genes if the reciprocally re-annotated genomes are used as mapping references. The reciprocal re-annotation pipeline can easily be adopted to re-annotate genomes of other closely related species or populations of animals and plants. Although we introduced our novel approach here to re-annotate three genomes at a time, it can of course be applied to two or more genomes.

4.2.5 Materials and Methods

4.2.5.1 Comparison of published annotations

We obtained the complete coding sequence (CDS) set of *D. melanogaster* r5.55 from FlyBase and considered only the longest isoform of each gene. Because identical sequences cannot be distinguished when RNA-seq reads are mapped (e.g. 23 different Histone 3 loci), we only retained one copy of genes with exactly the same nucleotide sequence (49 sequences, 195 transcripts discarded).

The genome and annotation of *D. mauritiana* was downloaded from http://www.popoolation.at/mauritiana_genome/index.html (Nolte et al., 2013), combining the five gene set files. The transcript set was obtained from a GFF file and the *D. mauritiana* genome. IDs were converted with the FlyBase conversion tool.

The genome and annotation of *D. simulans* was downloaded from http://genomics.princeton.edu/AndolfattoLab/w501_genome.html (Hu et al., 2013), combining “clean” and “unclean” data sets. The transcript set was obtained from a GFF file and the *D. simulans* genome.

Common genes were identified by gene ID (FBgn nomenclature) correspondence in all species. Genes absent from these species-specific annotations were identified by comparing the annotated genes to the genes present in the *D. melanogaster* gene set (data not shown). The absence of these genes was confirmed by tblastn (Altschul et al., 1990) search.

4.2.5.2 Direct re-annotation of genomes

The *D. mauritiana* and *D. simulans* genomes were obtained as described above and annotated with the *D. melanogaster* CDS set using Exonerate v2.2 (Slater and Birney, 2005) with the command `--model est2genome --softmasktarget yes --bestn 1 --minintron 20 --maxintron 20000`. The resulting GFF files were converted into transcript sets for each species from the corresponding genome files.

For some genes these three species have a different number of paralogs. For differential expression analysis it is essential to only consider orthologs of each gene, i.e. the number of reads that map to one transcript in one species cannot be reliably compared to the number of counts in two or more transcripts in another species. To count the total number of recovered transcripts in each annotation round, we kept only one copy of transcript sequences that gave more than one best hit in the target set. We selected the copy to keep based on conserved synteny (the putative paralog that is in the same chromosome and relative strand in the target genome and that has the same neighboring genes as in *D. melanogaster*) and conserved gene structure (the putative paralog that has the same number of exons as *D. melanogaster*). Genes for which none of the multiple copies found satisfied these conditions were discarded. In the *D. mauritiana* direct re-annotation only one gene gave more than one predicted copy (FBgn0264343); since none of the copies was in the same chromosome as *D. melanogaster* (2L) they were discarded. In the *D. simulans* direct re-annotation five genes gave more than one copy (FBgn0002933, FBgn0010294, FBgn0036177, FBgn0053874 and FBgn0062565); for the first three genes, the copy that was in the same relative strand as *D. melanogaster* was kept, FBgn0053874 was discarded because none of the copies was in the same chromosome as *D. melanogaster* (2L) and for FBgn0062565 only the copy predicted in the same chromosome as *D. melanogaster* (X) and with the same number of exons (3) was kept and the other was discarded.

BLAST 2.2.26+ (Altschul et al., 1990) was used to back-blast the resulting gene sets to the *D. melanogaster* gene set (`blastn -max_target_seqs 1`). Only the genes that had as best hit the *D. melanogaster* gene that had been used to annotate them (reciprocal best hit) were kept and reported in Table 4.2.1.

4.2.5.3 Generation of comparable transcriptomes – Reciprocal re-annotation pipeline

To generate reference transcriptomes for the three species with a minimum length difference between orthologous sequences and including the maximum number of

transcripts present in all species for analysis of inter-specific differential expression, we annotated the transcript sets of the different species via multiple rounds of pair-wise alignment with Exonerate v2.2 (Slater and Birney, 2005) following the scheme shown in Figure 4.2.5. Since FlyBase (St. Pierre et al., 2014) maintains an up to date curation and annotation the of *D. melanogaster* genome, we used this gene set as the first reference.

We used the *D. melanogaster* CDS set (r5.55) to annotate the *D. simulans* reference genome (Figure 4.2.5, step 1) with `exonerate --model est2genome --softmasktarget yes --bestn 1 --minintron 20 --maxintron 20000`. The resulting gene set was used to annotate the *D. mauritiana* reference genome using `--model est2genome` (Figure 4.2.5, step 2). At this point, the transcript set contains the maximized number of comparable genes and minimized transcript length difference between the three species' references. Consequently, step 3 consisted of reciprocally annotating the *D. simulans* transcript set with the *D. mauritiana* transcript set (Figure 4.2.5, step 3) and finally using the resulting *D. simulans* transcript set to annotate *D. melanogaster* transcript set (Figure 4.2.5, step 4). The criteria used to deal with multiple paralogs was the same as described above when the annotation reference was a genome (steps 1 and 2). Step 1 was the same as previously described and only one copy of FBgn0002933, FBgn0010294, FBgn0036177 and FBgn0062565 were kept. In step 2, only one gene (FBgn0263247) gave two hits in *D. mauritiana*; these two were clear tandem duplicates and the one predicted at 3L:11061688-11061810 was kept. In steps 3 and 4 only the genes where the gene ID of the target and the query matched were kept.

A back-blast to the original *D. melanogaster* gene set was also performed with the resulting gene sets of the three species. Only the reciprocal best hits were kept and reported in Table 4.2.1.

A list of gene names (FBgn nomenclature) and the respective transcript lengths for all annotations used in this study (published annotations, direct re-annotation and the reciprocal re-annotation) of all three species are available as part of the processed files uploaded to the Gene Expression Omnibus (GEO) database (Accession number: GSE76252). Additionally, gff and fasta files of the final datasets and of intermediate steps of the reciprocal re-annotation pipeline are available from GSE76252 as well.

4.2.5.4 RNA isolation and sequencing

RNA-seq reads for analysis of differential expression were generated for *D. melanogaster* (OregonR), *D. mauritiana* (TAM16, collected in Mauritius in 2007 (Nolte et al., 2013)) and

D. simulans (*yellow vermillion forked*, YVF; DSSC, University of California, San Diego, Stock no.14021-0251.146). In summary, flies were raised at 25°C and 12h:12h dark/light cycle in density-controlled conditions (30 freshly hatched LI larvae per vial). Female LIII larvae were dissected and eye-antennal imaginal discs were stored in RNALater (Qiagen, Venlo, Netherlands) at 120 h after egg laying. We dissected 40-50 discs per sample and generated three biological replicates for *D. melanogaster* and for *D. simulans* and 6 biological replicates for *D. mauritiana* (total of 12 samples).

Total RNA was isolated using the Trizol (Invitrogen, Thermo Fisher Scientific, Waltham, Massachusetts, USA) method according to the manufacturer's recommendations and the samples were DNase I (Sigma, St. Louis, Missouri, USA) treated in order to remove DNA contamination. RNA quality was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) microfluidic electrophoresis. Only samples with comparable RNA integrity numbers were selected for sequencing.

Library preparation for RNA-Seq was performed using the TruSeq RNA Sample Preparation Kit (Illumina, catalog ID RS-122-2002) starting from 500 ng of total RNA. Accurate quantitation of cDNA libraries was performed by using the QuantiFluor™dsDNA System (Promega, Madison, Wisconsin, USA). The size range of final cDNA libraries was determined applying the DNA 1000 chip on the Bioanalyzer 2100 from Agilent (280 bp). cDNA libraries were amplified and sequenced by using cBot and HiSeq 2000 (Illumina): single-end reads were generated for *D. mauritiana* (replicates A, B and C) and for *D. melanogaster* samples (1x50 bp) and paired-end reads were generated for *D. mauritiana* (replicates D, E and F) and for *D. simulans* samples (2x100 bp).

Sequence images were transformed to bcl files using the software BaseCaller (Illumina). The bcl files were demultiplexed to fastq files with CASAVA (version 1.8.2). Quality control was carried out using FastQC (version 0.10.1, Babraham Bioinformatics). Only replicates A, D and E from *D. mauritiana* and replicate C from *D. simulans* had bases with Phred quality score <Q20. Following recently published guidelines (Macmanes, 2014) we did not trim these bases but instead relied on the aligner software to make the quality call. Due to this procedure the overall mapping success (% mapped reads) for all datasets was slightly reduced. Of *D. melanogaster* (replicate A) for example, about 4.8% of the reads do not map against the entire genome, suggesting that they might be filtered out due to low quality during the mapping procedure (Supplementary Table 3).

Raw fastq files of all samples have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE76252 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76252>).

4.2.5.5 Analysis of differential expression

Since we generated two different types of RNA-seq reads (namely 100 bp paired-end and 50 bp single-end), we only compared the datasets that were produced with the same technique, i.e. *D. melanogaster* reads were compared only to *D. mauritiana* 50 bp reads and *D. simulans* reads to *D. mauritiana* 100 bp paired-end reads. Since 50 bp single-end reads are informative enough for differential expression analysis (Chhangawala et al., 2015; González and Joly, 2013; Li and Dewey, 2011) and this is the cutoff we set in our analysis as the maximum gene length difference, prior to mapping, 100 bp paired-end reads from *D. simulans* and *D. mauritiana* were split into two 50 bp reads each. Left and right reads were merged into a single file to be equivalent to single-end reads. 50 bp single-end reads from *D. mauritiana* and *D. melanogaster* were not processed prior to mapping.

Bowtie2 (Langmead and Salzberg, 2012) with parameters `-very-sensitive-local -N 1` was used in all cases to map the reads to the respective references: *D. melanogaster* reads were mapped to the published gene set (Flybase, r5.55) and to our novel reciprocally re-annotated gene set. *D. mauritiana* and *D. simulans* reads were mapped to the respective published gene sets (Hu et al., 2013; Nolte et al., 2013), to the directly re-annotated gene sets and to the reciprocally re-annotated gene sets. The number of reads mapping to each transcript were summarized using samtools v0.1.19 (Li et al., 2009).

To calculate the percentage of reads mapped to UTRs we aligned *D. melanogaster* replicate A reads to the longest full transcripts of *D. melanogaster* r5.55 and compared the mapping percentage to that of the mapped reads to the longest CDS set. To calculate the percentage of reads mapped to transcript regions not included in the longest CDS set we aligned *D. melanogaster* replicate A reads to the complete CDS set (including all isoforms) and compared the mapping percentage to that of the longest CDS set. To calculate the percentage of reads generated from unannotated regions we aligned *D. melanogaster* replicate A reads to the complete *D. melanogaster* genome r5.55 and compared the mapping percentage to that of the mapped reads to all annotated transcripts.

Differential expression was determined for each orthologous gene between *D. melanogaster* and *D. mauritiana* (from the originally 50 bp single-end reads) and between *D. simulans* and

D. mauritiana (from the originally 100 bp paired-end reads). Four different methods were used to call differentially expressed genes for each annotation strategy:

1. DESeq2 (Love et al., 2014b) (v1.6.3) with direct counts per transcript and default parameters.
2. DESeq2 with a transcript length normalization factor matrix with row-wise geometric means of 1. This matrix was applied with the command `normalizationFactors()`. The rest of parameters were left as default.
3. Limma (Ritchie et al., 2015; Smyth, 2004) (v3.22.7) on reads per kilobase per million (RPKM). RPKM values were calculated for each transcript with the corresponding library size and transcript length. 1 was added to the resulting value to prevent negative values when applying log transformation. Limma was applied to \log_2 transformed RPKM values to call differentially expressed genes using `ebayes(trend=T)`.
4. RPKM-voom-limma (Law et al., 2014). RPKM values were calculated as described above and `voom()` was used with default parameters to log-transform the data and obtain the associated precision weights matrix. Limma with default parameters was applied to the resulting `ELIST` object to perform the differential expression analysis.

For all methods, Benjamini & Hochberg correction was used to adjust p-values for multiple testing (default in DESeq2 and Limma). Genes were called significantly differentially expressed when the program reported an adjusted p-value lower than 0.05.

R (v3.1.2) (R Core Team, 2015) was used to generate the correlation plots. The Venn diagrams were generated using `jvenn` (Bardou et al., 2014). IGV (v2.3) (Robinson et al., 2011; Thorvaldsdottir et al., 2013) was used to visualize read coverage of the *Cp110* transcript and Mafft (v7.017) (Katoh et al., 2002) (as integrated in Geneious v6.0.6 (Biomatters, Auckland, New Zealand)) was used to align the annotated *Cp110* transcripts of *D. melanogaster* and *D. mauritiana*.

4.2.5.6 Real-time qPCR

RNA from eye-antennal imaginal discs from female LIII larvae was extracted using ZR Tissue & Insect RNA MicroPrep™ (Zymo Research, Irvine, CA, USA). RNA concentration was measured using Qubit (Invitrogen, Thermo Scientific, Waltham, Massachusetts, USA). Samples were diluted to contain exactly the same amount of starting RNA. RNA was converted to cDNA using MAXIMA® First Strand cDNA synthesis for RTqPCR (Thermo Scientific, Waltham, Massachusetts, USA). For the “no RT” control

parallel reactions were carried out without enzyme. For the efficiency test, a series of five 1:4 dilutions were made. Real-Time qPCR was performed with HOT FIREpol® EvaGreen® qPCR Mix Plus (ROX) (Solis BioDyne, Tartu, Estland) in a CFX96™ Real-Time PCR System (Bio-Rad Laboratories, Hercules, CA, USA). Primers were designed to exclude polymorphisms between *D. melanogaster* (FlyBase) and *D. mauritiana* TAM16 and to amplify a sequence that span introns to avoid genomic contamination (except for *Cp110*, *alrm* and *actin 79B*) and did not show isoform variation. Primer sequences are given in Supplementary Table 4. A melting curve was performed at the end of each reaction. Only genes that produced a single peak are shown. Expression differences were calculated by $\log_2(2^{-\Delta\Delta C_t})$, using *actin 79B* as reference gene. Differences in expression were assessed using t-test/t-Welch-test with FDR=0.05.

4.3 Gene expression divergence in closely related *Drosophila* species

In the first project of this Thesis, “**New regulatory interactions governing *Drosophila* head development**”, I identified the genes that are differentially expressed during the development of the larval eye-antennal imaginal discs, and grouped these genes according to their expression profile during head and eye development in the model species *D. melanogaster*. The closely related *Drosophila* species *D. melanogaster*, *D. mauritiana* and *D. simulans* show clear differences in head morphology (Arif et al., 2013; Hilbrant et al., 2014; Posnien et al., 2012). Thus they are a good model to study the mechanisms by which natural selection has allowed morphological differences to evolve while keeping functioning developmental GRNs. Initially, I wanted to investigate if the expression dynamics throughout eye-antennal imaginal disc development are the same in other closely related species. Thus, by identifying the genes that have conserved expression in different species, I can obtain the genes that represent the core players of this biological process. Afterwards, by investigating the genome-wide differences in gene expression between these species, I will first reveal all differentially expressed genes, and later also identify the mechanisms by which this divergence in gene expression is regulated.

4.3.1 Developmental transcriptome of three closely related *Drosophila* species

4.3.1.1 Evaluation of bias introduced by the use of different sequencing types

I sequenced the transcriptomes of eye-antennal imaginal discs of the closely related species *D. mauritiana* TAM16 and *D. simulans* YVF at the same larval stages as I had information for *D. melanogaster* (late LII, mid LIII and late LIII larvae) (Table 3.2).

It is relevant to note that all these samples were generated at different times with different sequencing types (i.e. 100 bp paired-end reads vs. 50 bp single-end reads) (see Table 3.2 and Materials and Methods for details). In order to exclude batch effects caused by different sequencing time points and sequencing types, I sequenced one of the samples using the two sequencing types (*D. mauritiana* eye-antennal imaginal discs at 120h AEL) and performed thorough quality tests.

To investigate whether the use of different sequencing types could introduce a bias in the data, I performed multidimensional scaling clustering of all the samples (Figure 4.3.1). In

the first panel, the samples are colored by stages, and it is clear that this is the factor accounting for the biggest difference between samples, since samples from 72h appear far apart from the rest (Figure 4.3.1A). The second component separates the data by species (Figure 4.3.1B), since the *D. melanogaster* samples are on the top part of the plot and the other two species (which are more closely related) are at the bottom. The six *D. mauritiana* 120h replicates cluster together, regardless of the sequencing type. All replicates are equally separated by the first component (dimension 1) and appear close to each other and to the 0 when separated by the second component. This indicates that the use of different sequencing types did not introduce a clear bias in the data.

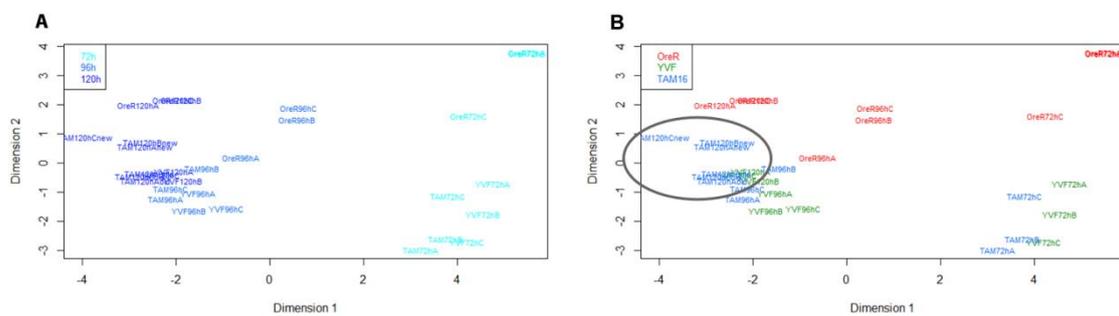


Figure 4.3.1. Multidimensional scaling plots of three species' samples. (A) Samples are colored according to the different time points: 72h, 96h or 120h AEL. This condition corresponds to the first dimension separating the data. **(B)** Samples are colored according to the different species: *D. melanogaster* OreR, *D. simulans* YVF or *D. mauritiana* TAM16. This condition corresponds to the second dimension separating the data. *D. mauritiana* TAM16 120h AEL samples are indicated with a grey circle, showing that the use of different sequencing types does not greatly influence the separation of the samples.

4.3.1.2 Conserved gene expression during *Drosophila* eye development

Once I confirmed that the data were not biased due to the different sequencing types, I used all the samples to identify the genes whose expression changes during eye-antennal imaginal disc development in the three *Drosophila* species. I aimed to study if the developmental processes and gene clusters that I had identified in *D. melanogaster* are the same in these sister species, and by that be able to identify the genes that have a conserved expression throughout development. I performed an analysis with HTScluster (Rau et al., 2015) analogous to the one performed in section 4.1.2, but including all count data for the three species and the three stages. This initially resulted in 8 clusters (DDSE model), but I noted that 6 of these clusters (clusters 3 to 8) were very similar and included only genes with high expression at 72h. Clusters 1 and 2 contained 2,956 and 1,011 genes, respectively, and included all genes that were up-regulated in later stages. To get a better resolution of the changes in expression profile of these genes, I re-clustered the genes in clusters 1 and 2.

This resulted in 7 more subclusters (DDSE and Djump gave the same result), and therefore a total of 13 clusters (Figure 4.3.2).

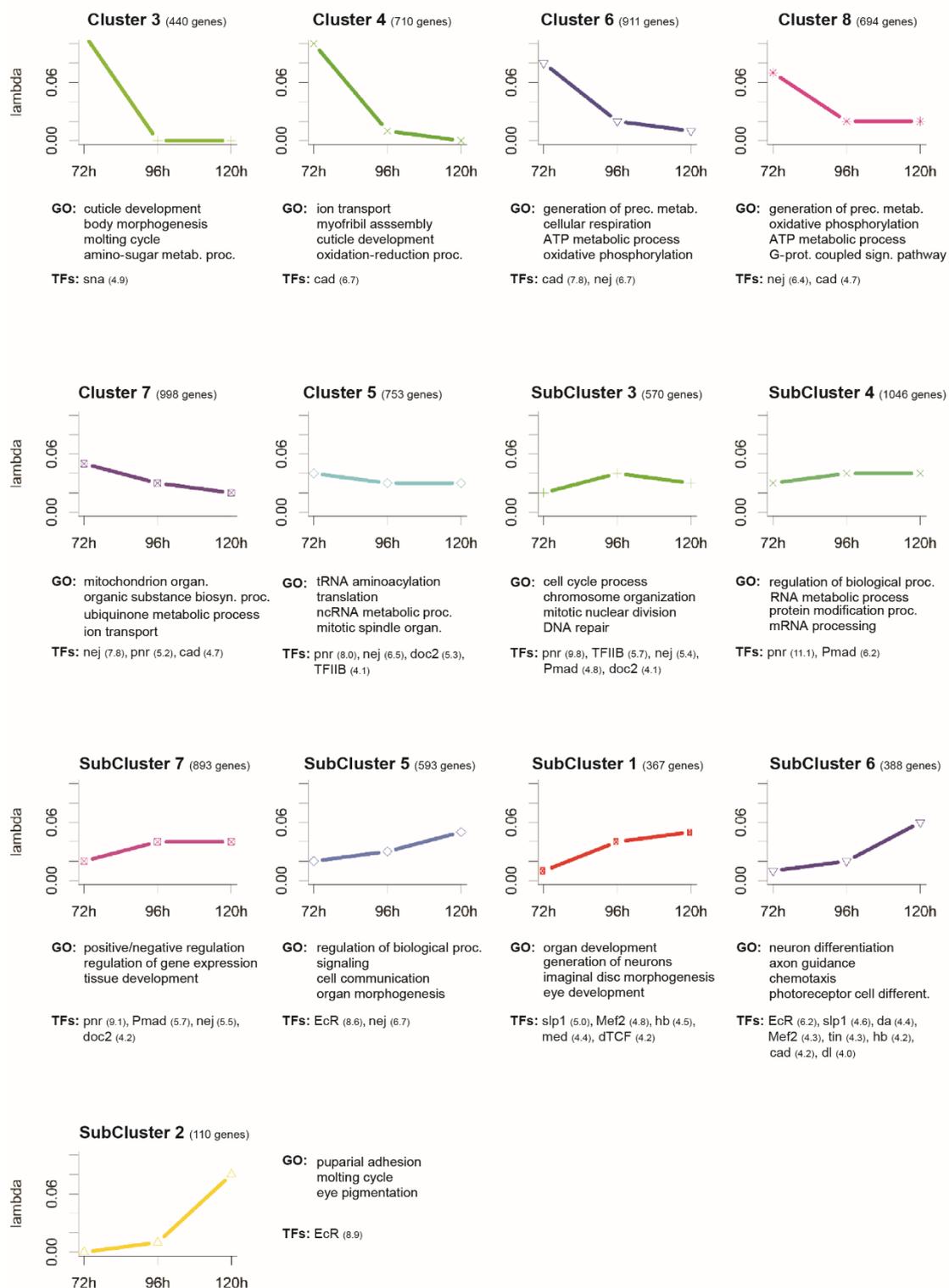


Figure 4.3.2. Co-expression clusters in three *Drosophila* species. Lambda values for each stage and each cluster. Under each cluster the first 4 non-redundant enriched GO terms are listed and the enriched transcription factors (NES ≥ 4.0) predicted by *i-cisTarget*. Normalized enrichment score (NES) is indicated in brackets.

A comparison of the obtained clusters in this analysis and that of only *D. melanogaster* reveals that the identified expression profiles are very similar in both analyses (compare Figure 4.3.2 to Figure 4.1.3). The GO terms enriched in each cluster were also similar to those obtained with only one species (Figure 4.3.2; see Discussion 5.1.1). Finally, I also identified the transcription factors that are likely to regulate a large number of co-expressed genes of each of these 13 clusters using i-cisTarget (Herrmann et al., 2012) (Figure 4.3.2). As in the analysis using only *D. melanogaster* developmental transcriptome, Nejire, Pannier and Caudal are enriched as possible regulators of the genes present in clusters of genes expressed at the early stage when I use the data for the three closely related species. Additionally, other transcription factors like Slp1, Mef2 and Hb appear as well as putative regulators of the genes in two clusters with genes up-regulated at late stages.

4.3.1.3 Inter-species differential gene expression

After finding out that there is a great conservation of gene expression dynamics between these three species during eye-antennal imaginal disc development, I set out to investigate which genes have divergent gene expression. First I carried out a pair-wise differential expression analysis between each pair of species for each time point (Figure 4.3.3). The largest differences can be observed between *D. melanogaster* and the other two species at 72h, when 5,097 genes are differentially expressed between *D. melanogaster* and *D. simulans* and 6,032 genes are differently expressed between *D. melanogaster* and *D. mauritiana*. In contrast, only 697 genes have different expression levels at 72h between the more closely related species *D. simulans* and *D. mauritiana*. At 96h this tendency is the same, and there are more genes differentially expressed between *D. melanogaster* and the other two species than between these two species with each other. Interestingly, between *D. melanogaster* and *D. simulans*/*D. mauritiana* there are less differentially expressed genes than at 72h, while between *D. simulans* and *D. mauritiana* the number of differentially expressed genes is larger at 96h than at 72h. Finally, at 120h there are even more genes with differential expression between *D. simulans* and *D. mauritiana*. The number of differentially expressed genes between *D. melanogaster* and *D. mauritiana* at this stage is similar to that at 96h. However, between *D. melanogaster* and *D. simulans* the number of differentially expressed genes is very high.

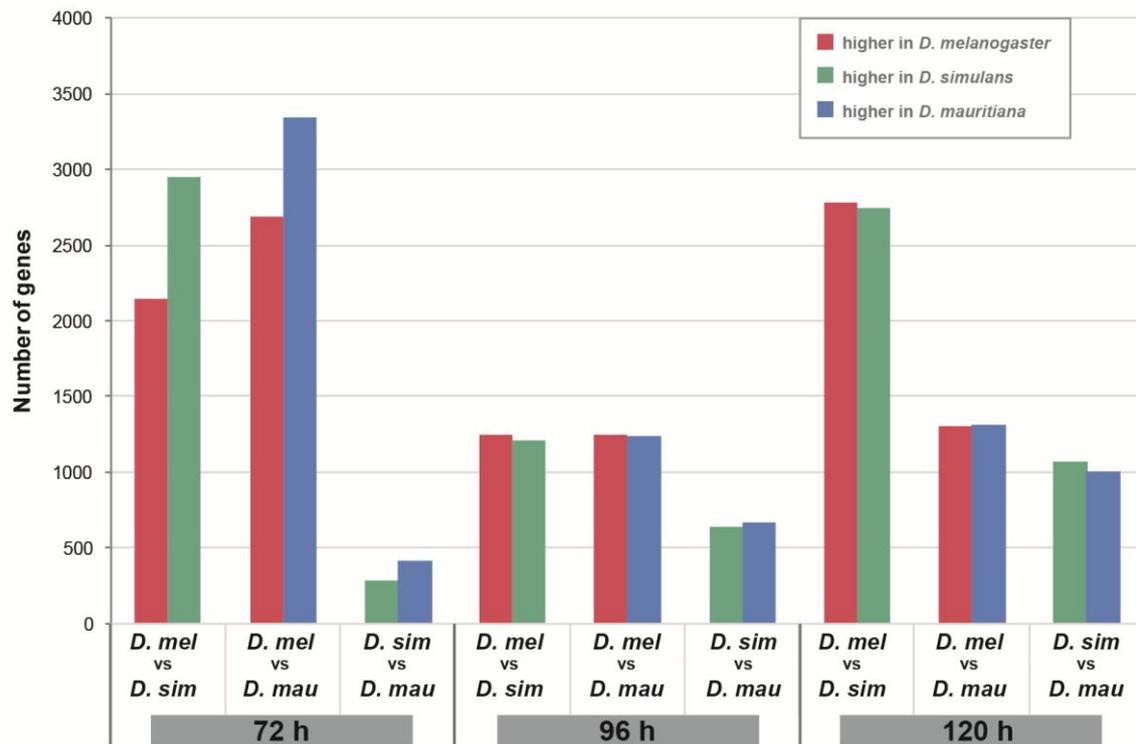


Figure 4.3.3. Pair-wise differential inter-species gene expression. Bar plot showing the number of genes differentially expressed in each pair-wise comparison.

To obtain a better picture of the groups of genes that change their expression at each time point and species, first I performed a multi-factor DESeq2 (Love et al., 2014a) analysis to identify the genes differentially expressed across all samples. I set the analysis parameters to identify the genes that varied the most between the different species, minimizing the variation caused by the different stages. A total of 6,649 genes appeared as significantly differentially expressed between *D. melanogaster* and *D. simulans* ($p\text{-adj} < 0.05$) when considering the three developmental stages.

The heat map showing the hierarchical clustering of the 1,000 most differentially expressed genes (lowest p-adjusted values) shows that the largest differences in expression can be observed between *D. melanogaster* and the other two species, although the expression across stages is not homogeneous (Figure 4.3.4). Even though the variation between stages was minimized, a clear difference between gene expression at 72h and the later stages can be observed. Interestingly, when grouping these genes into different clusters, some groups with similar expression profiles can be identified by the distinct predicted clusters (Figure 4.3.4). About 30% of genes have high expression across all samples but at 72h they have significant differential expression in *D. simulans* and *D. mauritiana* compared to *D. melanogaster* (blue cluster). Based on an i-cisTarget analysis (Herrmann et al., 2012) these

genes are likely to be regulated by the transcription factors Pannier, Pmad and Dorsocross2. The orange cluster contains genes that in all *D. simulans* and *D. mauritiana* samples have higher expression than in *D. melanogaster*, and these genes could be regulated by Caudal, Mef2 and Twist. In contrast, the green cluster contains genes with higher expression in all *D. melanogaster* samples compared to the other two species, and these genes, apart from being enriched to be putatively regulated by Caudal, present the distinctive GATA binding motif in their regulatory regions.

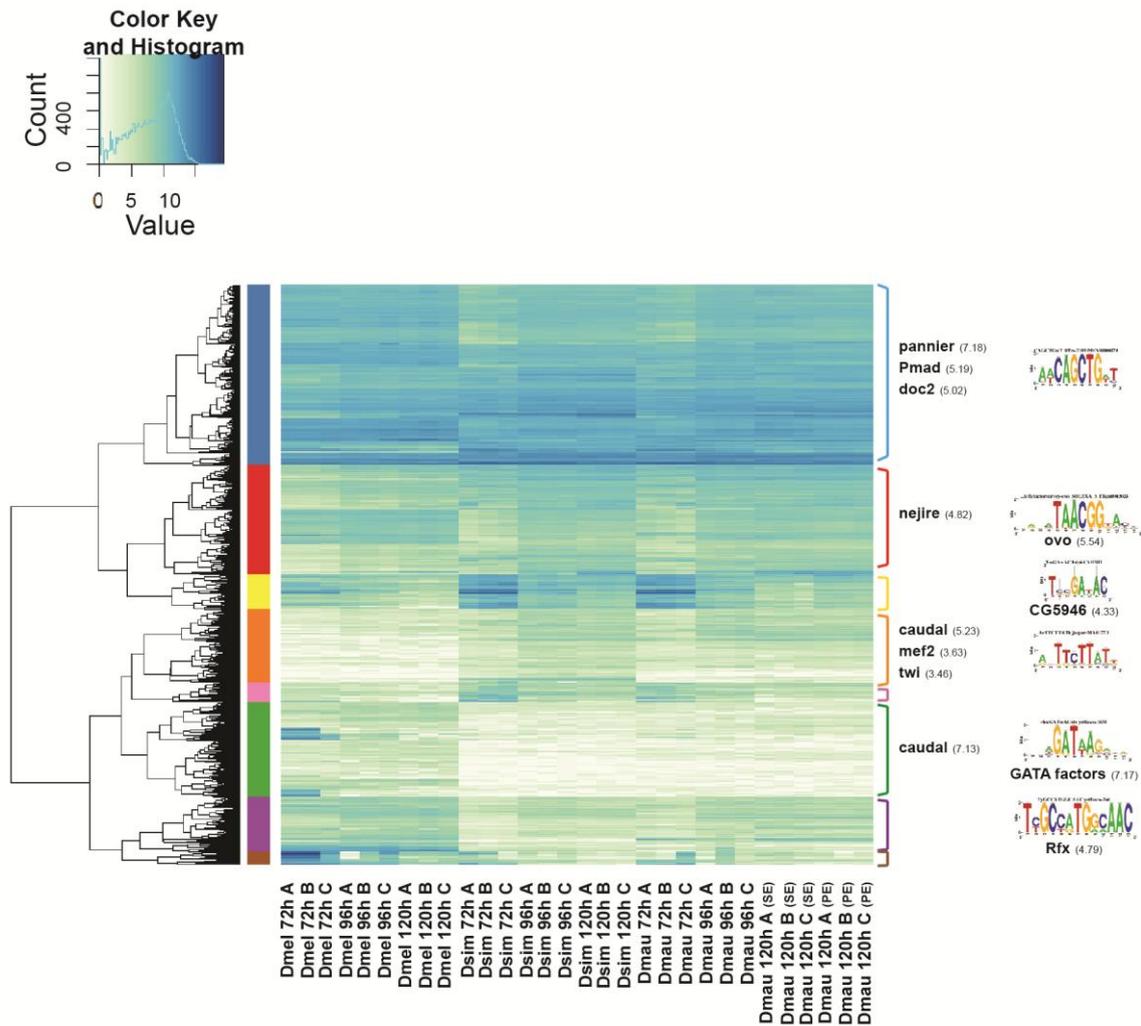


Figure 4.3.4. Heat map of expression differences between *Drosophila* species. Heat map representing the expression of the 1,000 genes that are most differentially expressed between species (*D. melanogaster* vs. *D. simulans*). Each row in the heat map represents one gene and the color in each cell (from white to dark blue) represents the normalized expression level as indicated in the color key (top left corner). Genes are ordered by hierarchical clustering based on the distances dendrogram (left side) and grouped into 8 clusters according to their expression profile (different vertical colored bars between the dendrogram and the heat map). Samples are indicated at the bottom of the heat map. On the right side of the heat map are listed the enriched transcription factors with NES ≥ 3.0 . Next to that the highest scoring PWM is shown; when known, the corresponding transcription factor is given below the motif. The NES score representing a confidence level given by i-cisTarget is indicated in brackets for each transcription factor or PWM.

4.3.2 Evolution of gene expression differences

A recurrent question in evolutionary biology is the influence of network topology on gene expression divergence (Carlson et al., 2006; Siegal et al., 2007; Ulitsky and Shamir, 2007). For instance, are more central factors with likely stronger pleiotropic effects prone to show expression differences or are more changes observed in genes with fewer connections (e.g. terminal genes with less pleiotropic functions)? Therefore, after identifying the differentially expressed genes between these three closely related *Drosophila* species, I wanted to know where these divergent genes are located in the molecular networks involved in eye and head development. For that I mapped the inter-species differential gene expression data on the networks of genetic interactions generated from the clustering of *D. melanogaster* developmental transcriptomic data (section 4.1.2) (Figure 4.3.5). In cluster 3 only one gene with more than 3 interactions (*asp*) is differentially expressed (Figure 4.3.5A). In cluster 10 one gene with 10 genetic interactions (*Nsf2*) is differentially expressed, and the genes from the small interconnected network *kay-puc-slpr* are all significantly higher expressed in *D. mauritiana* compared to *D. melanogaster* (Figure 4.3.5B). Interestingly, all genes differentially expressed in cluster 11 have higher expression in *D. mauritiana* at 96h AEL, including highly interconnected genes such as *ss*, *aop*, *syp*, *hb*, *Ret*, *Dl*, *Abl*, *ena*, *sty* and *hid* (Figure 4.3.5C).

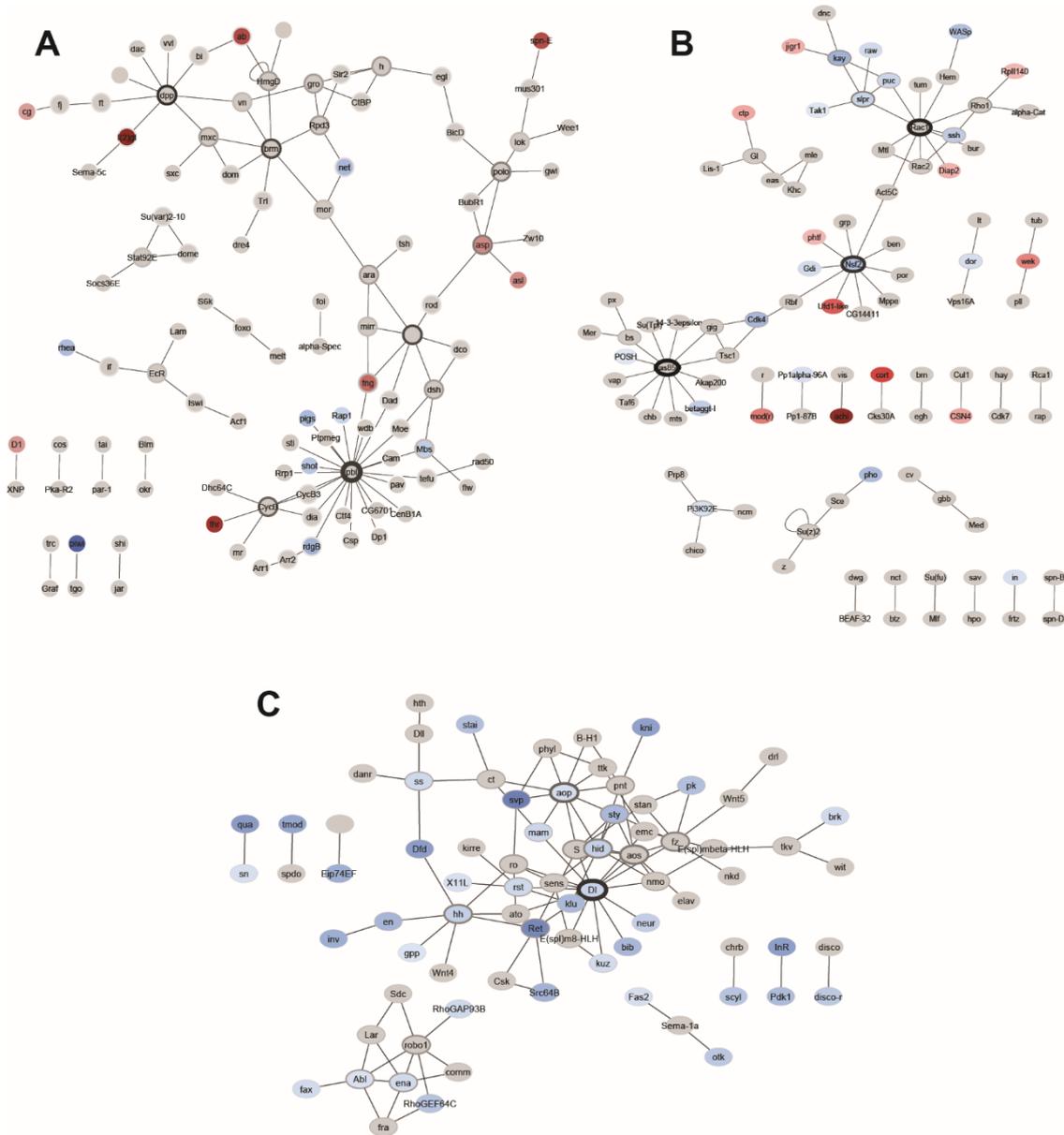


Figure 4.3.5. Differentially expressed genes in the genetic interaction networks. Genes are represented as nodes and genetic interactions as edges. Blue shaded circles indicate genes with higher expression in *D. mauritiana* at 96h AEL, while red shaded circles indicate higher expressed genes in *D. melanogaster* at the same stage. Darker shade indicates higher fold-change. The circle margin is thicker for genes with more interactions. **(A)** Cluster 3. Nine genes have higher expression in *D. melanogaster* and 8 are higher in *D. mauritiana*, where 5 of these interact with the gene *pbl*. **(B)** Cluster 10. 11 genes have higher expression in *D. melanogaster*, although none of them have known connections with each other, and 17 are higher in *D. mauritiana*, with a whole cluster of 5 genes around *spr*. **(C)** Cluster 11. All genes differentially expressed in this cluster are higher in *D. mauritiana*.

4.3.3 Detection of *cis* and *trans* regulatory divergence by allele-specific expression (ASE) analysis

Recent studies have used high-throughput transcriptomic data of F₁ hybrid organisms to study the relative contribution of *cis* and *trans* variation to the generation of gene expression divergence between closely related species (e.g. Graze et al., 2012; Tirosch et al., 2009; Zhang and Borevitz, 2009). These studies are based on the comparison of allele-specific expression in the hybrid individuals to the relative gene expression in their parents (see also Figure 2.3). We took advantage of the viability of F₁ hybrid individuals between the *Drosophila* species and we crossed *D. melanogaster* females with *D. mauritiana* males and *D. simulans* females with *D. mauritiana* males. For each cross we sequenced the transcriptomes of F₁ hybrids for eye-antennal imaginal discs (96h AEL and 120h AEL) and for wing imaginal discs (96h AEL).

4.3.3.1 Generation of polymorphism-rich strain-specific references and allele-specific read mapping

An important prerequisite for the analysis of ASE is the presence of polymorphisms in the parental species that allow the distinction of the species of origin of the hybrid reads (Wittkopp et al., 2004). RNA-seq technology generates only reads of short length, the larger the frequency of polymorphisms between the pairs of orthologs, the more reads can be mapped and used to analyze gene expression divergence. However, the closer the two species are related phylogenetically, the fewer number of polymorphisms exist between them. Thus, the bioinformatics analyses required for these studies is especially challenging and some steps need to be taken to prepare the references prior to mapping the reads (Stevenson et al., 2013).

The transcriptome references used for inter-species differential expression analyses in the previous sections (Torres-Oliva et al. in revision; section 4.2) were based on the re-annotation of the previously published genomes of *D. melanogaster* (Hoskins et al., 2007), *D. mauritiana* (Nolte et al., 2013) and *D. simulans* (Hu et al., 2013), but in all cases the strain used was different from the one I have used in my analyses (*D. melanogaster* iso-1, *D. mauritiana* MS17 and *D. simulans* w⁵⁰¹ are published). Additionally, in Torres-Oliva et al. (in revision) I only used the coding sequences to perform the reciprocal re-annotation of these genomes. Therefore, I first examined how many polymorphisms existed between the existing references for the species for which hybrid data was available, i.e. between *D. mauritiana* and *D. melanogaster* (Figure 4.3.6A and Table 4.3.1) and between *D. mauritiana* and

D. simulans (Figure 4.3.6B and Table 4.3.1). Concordant with these species' phylogeny, the number of polymorphisms between *D. melanogaster* and *D. mauritiana* is much larger than between *D. simulans* and *D. mauritiana*. Using only the coding sequence of the transcripts, very few genes have more than 50 mismatches between the more closely related species *D. simulans* and *D. mauritiana*, almost 2,000 genes have less than 5 mismatches and 231 genes have no mismatch that can differentiate the orthologous sequences. Between *D. melanogaster* and *D. mauritiana* there are only 39 genes without mismatches and 371 genes with less than 5 mismatches; however, most of the orthologous genes have less than 30 mismatches (Figure 4.3.6A).

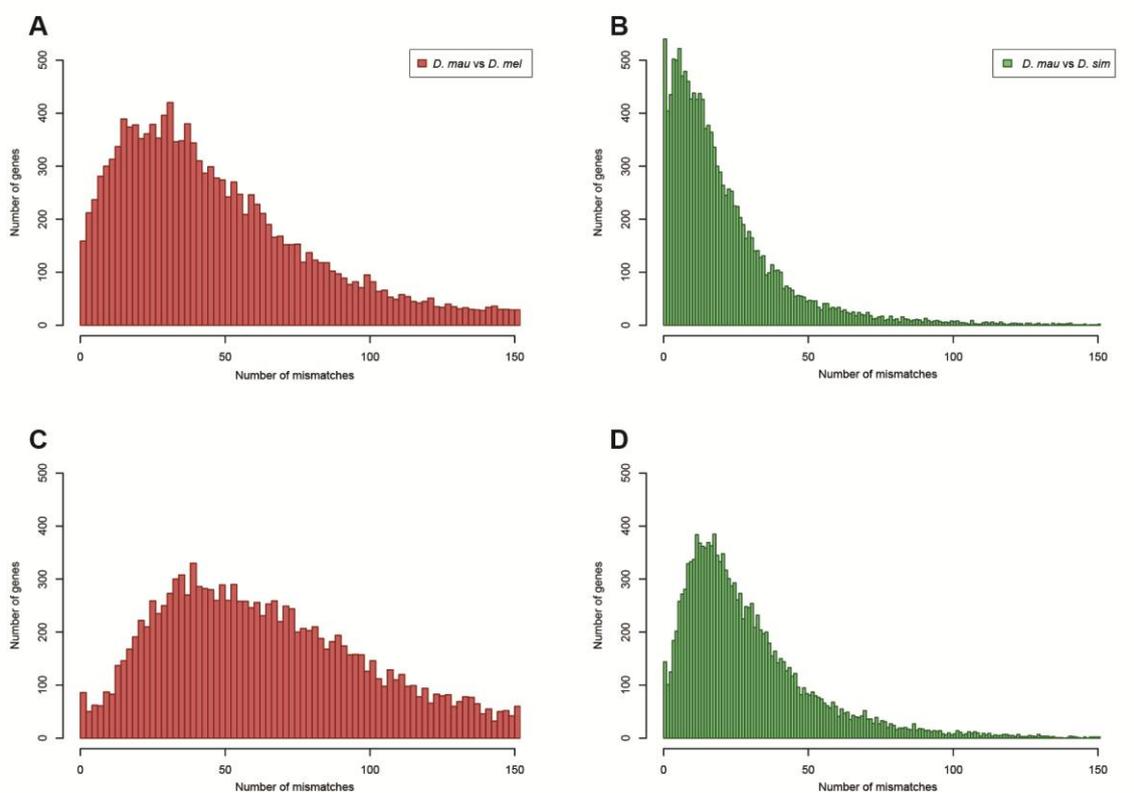


Figure 4.3.6. Mismatches between species references. Histogram of the number of genes presenting the specified number of mismatches between orthologs in the different annotated references. **(A)** Comparison of *D. mauritiana* and *D. melanogaster* orthologs annotated in the published genomes without UTR regions. **(B)** Comparison of *D. mauritiana* and *D. simulans* orthologs annotated in the published genomes without UTR regions and additional RNA-seq-based, strain-specific sequence replacement. **(C)** Comparison of *D. mauritiana* and *D. melanogaster* orthologs annotated in the strain specific genomes containing UTR regions. The peak is moved to the right compared to (A), fewer genes have less than 20 mismatches and there is also an increase in the number of genes with more than 70 mismatches. **(D)** Comparison of *D. mauritiana* and *D. simulans* orthologs annotated in the strain specific genomes containing UTR regions. The peak is slightly moved to the right compared to (B) and the number of genes with less than 15 mismatches is greatly reduced.

Table 4.3.1. Mismatches (mm) between orthologs in different references.

	<i>D. mau</i> vs <i>D. mel</i>		<i>D. mau</i> vs <i>D. sim</i>	
	only CDS	UTR	only CDS	UTR
0 mm	39	35	231	67
less than 5 mm	371	136	1881	526
1 mm/ 50 bp	12,431	13,321	3,146	3,938

These data showed that it was necessary to increase the number of detectable polymorphisms that could allow distinguishing the allele of origin of the hybrid RNA-seq reads. In order to do that, first I generated strain specific references by *in silico* polymorphism replacement at the genome level using strain specific genomic reads (see Methods). After that I repeated the reciprocal re-annotation pipeline between the two species pairs on these strain specific genomes, but this time using the full transcript sequences of *D. melanogaster* (including UTR) as starting reference. This strategy increased the number of mismatches per gene between the *D. mauritiana* and *D. melanogaster* orthologs (compare Figure 4.3.6A to Figure 4.3.6C) and almost all genes (13,321 genes) presented at least one mismatch per 50 bp of sequence (Table 4.3.1).

The number of mismatches between *D. mauritiana* and *D. simulans* orthologs also increased with this method (data not shown). However, preliminary analyses revealed that the *D. simulans* parental reads were not able to map to this species' reference using the very stringent parameters required to perform ASE studies (not shown; see Discussion 5.3.3.2). As a consequence, the hybrid reads mapped preferentially to the *D. mauritiana* allele and generated a great bias in the results. Thus I decided to perform a second round of *in silico* sequence replacement, this time using the species specific parental RNA-seq reads (see Methods). This strategy increased the number of mapped parental reads in *D. simulans* and reduced the bias in the mapped hybrid reads (not shown). Combined with the previously described methods, the number of mismatches between orthologs in *D. mauritiana* and *D. simulans* was increased (compare Figure 4.3.6B to Figure 4.3.6D), and only 67 genes had no recognizable polymorphisms (Table 4.3.1).

Once I obtained polymorphism-rich references that allowed a proper distinction of the allele of origin of the hybrid RNA-seq reads, I proceeded to map the parental and hybrid reads to the corresponding references. For each pair of species, the same parameters were used to map the parental reads to the species-specific reference and also the hybrid reads to the combination of the two species' references. Only those hybrid reads that mapped unambiguously to one allele were reported and counted. On average, around 65% of the

parental reads could be mapped to the species specific references with these stringent parameters (Table 4.3.2), with more than 30 million mapped reads in all samples. The mapping percentage of the hybrid samples was lower, since only reads that contained polymorphisms that allowed an unambiguous mapping were allowed. The percentage dropped for the *D. simulans* x *D. mauritiana* hybrids due to the less number of polymorphisms between the orthologs in these two species. Still, more than 10 million reads mapped in all replicates (Table 4.3.2).

Table 4.3.2. Mapping stats.

	tissue	percentage of mapped reads*	total mapped reads*
<i>D. melanogaster</i>	96h eye	72.26	34,436,404
	120h eye	72.18	30,599,797
	96h wing	73.71	43,959,111
<i>D. mauritiana</i>	96h eye	64.97	44,659,761
	120h eye	69.54	33,324,734
	96h wing	63.87	34,716,996
<i>D. simulans</i>	96h eye	62.99	33,098,213
	120h eye	62.42	49,206,491
	96h wing	61.04	35,986,926
<i>D. mel</i> x <i>D. mau</i>	96h eye	47.87	23,963,441
	120h eye	50.12	23,573,754
	96h wing	38.65	19,043,597
<i>D. sim</i> x <i>D. mau</i>	96h eye	21.20	11,822,104
	120h eye	21.80	14,899,185
	96h wing	20.16	10,921,398

*Mean of each triplicate of biological replicates.

4.3.3.2 Mitochondrial gene expression in F₁ hybrids

In most species, including *Drosophila*, mitochondrial DNA is only maternally transmitted (DeLuca and O'Farrell, 2012; Reilly and Thomas, 1980), thus all reads allocated to mitochondrial genes of hybrid animals should originate from the parental species that contributed as female in the cross. Therefore, the ASE of mitochondrial genes can be used as a control to check whether the expression of the mitochondrial genes in the hybrids originates from the female (*D. melanogaster* in the *D. mauritiana* x *D. melanogaster* cross and *D. simulans* in the *D. mauritiana* x *D. simulans* cross). In both analyses, practically all counted reads were from the species that contributed the female in the expressed mitochondrial genes (Figure 4.3.7). Only gene *mt:ND3* in the *D. mauritiana* x *D. simulans* cross had more reads recognized as the *D. mauritiana* allele. A closer inspection of the mapping in this gene

showed that the region where the hybrid reads mapped to the *D. mauritiana* allele contained a clearly unspecific base in the parental *D. simulans* reference (Supplementary Figure 11).

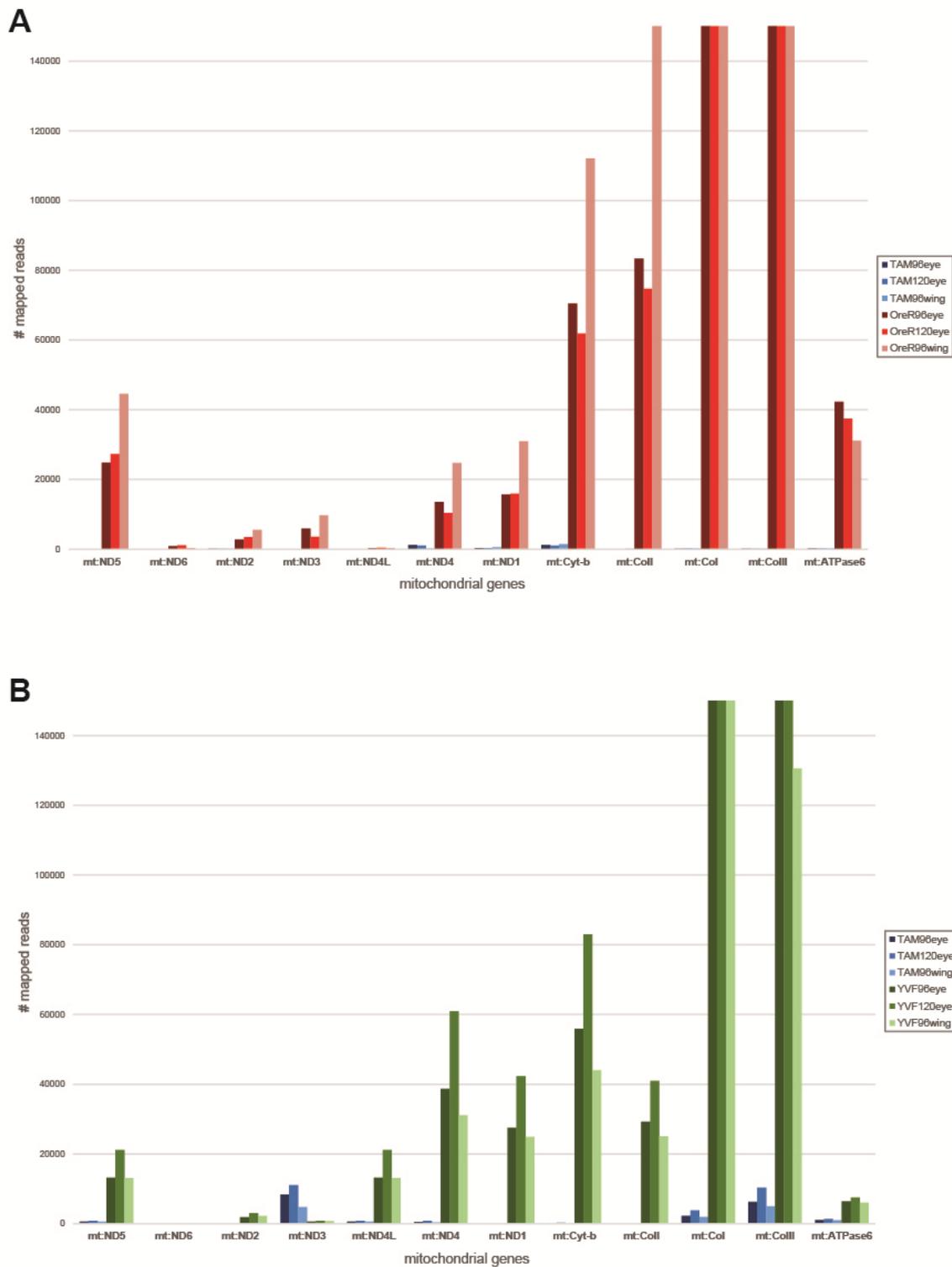


Figure 4.3.7. Allele-specific expression of mitochondrial genes in the hybrids. (A) *D. mau* x *D. mel* hybrids. *D. mauritiana* reads in shades of blue and *D. melanogaster* reads in shades of red. All genes present expression practically exclusive of the *D. melanogaster* allele, which is the mother in this hybrid cross. **(B)** *D. mau* x *D. sim* hybrids. *D. mauritiana* reads in shades of blue and *D. simulans*

reads in shades of green. All genes but ND3, CoI and CoIII present only expression of the *D. simulans* allele, which is the mother in this hybrid cross. Many more reads are identified as coming from the *D. simulans* allele than from the *D. mauritiana* allele in CoI and CoIII.

4.3.3.3 Gene expression differences are mainly caused by changes in *trans*

After confirming that almost all genes could be identified by polymorphisms present between orthologous genes (Figure 4.3.6) and that the strategy to detect allele-specific expression worked correctly (Figure 4.3.7), I proceeded to analyze the type of regulation that is responsible for the divergence in gene expression between the studied closely related species. First I used DESeq2 (Love et al., 2014a) to detect differentially expressed genes between the parental species (as described in section 4.3.1.3, Figure 4.3.3), and afterwards I used the same method to detect differentially expressed alleles in the hybrid data. The majority of genes had conserved expression across the species, as neither the orthologs were significantly differentially expressed between the parent species nor the alleles were significantly differentially expressed in the hybrids (Figure 4.3.8). As described in the Introduction (Figure 2.3), the relative differential gene expression in the parents compared to the hybrids was used to discern the type of regulatory changes (*cis* or *trans*) that cause divergence in gene expression between these species (see also Figure 2.3). In short, genes that are differentially expressed in the parental animals but show no significant differential expression in the hybrids are assumed to have divergent expression due to variation in *trans*. Genes with equal differential expression in the parents and in the hybrids are classified as to be divergent due to variation only in *cis*. In case that the alleles of a gene are differentially expressed in the hybrids but the gene is not differentially expressed between the parents, compensatory regulation is assumed to be acting. Finally, *cis* \times *trans* regulation is considered in genes that are differentially expressed in one direction in the parents and in the opposite direction in the hybrids.

In my study, clearly most of the genes with divergent gene expression between *D. mauritiana* and *D. melanogaster* are different because of variation in *trans*, both in the eye and the wing imaginal discs at the studied stages (Figure 4.3.8A). This is also the case between *D. mauritiana* and *D. simulans* eye-antennal imaginal discs, where even a larger percentage of genes appear to have divergent expression due to changes in *trans* (Figure 4.3.8B). However, in wing imaginal discs most genes have compensatory regulation. The number of genes with divergent expression because of variation in *cis* is quite low in all tissues and stages, especially between eye-antennal imaginal discs in the *D. mauritiana* \times *D. simulans* cross.

Finally, practically no genes have divergent expression due to an interaction of *cis* and *trans* regulatory differences (*cis* x *trans*).

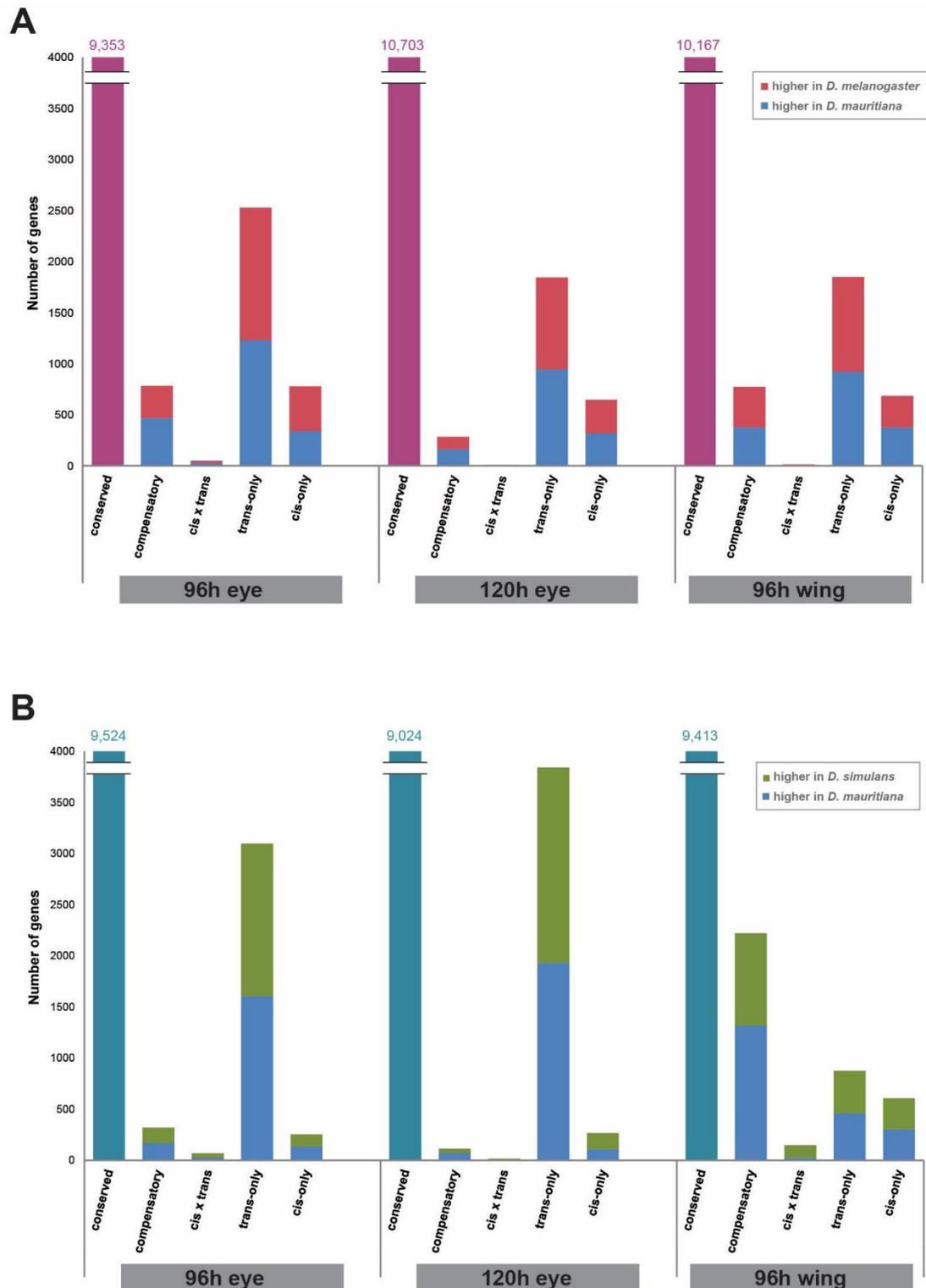


Figure 4.3.8. Regulation type. Classification of genes according to the type of regulatory changes that cause the difference in their relative expression in the parental species: only *cis*, only *trans*, *cis* x *trans* or compensatory. The first bar in each sample shows the number of genes with conserved expression. In grey background the sample is indicated: 96h AEL eye-antennal imaginal discs, 120h AEL eye-antennal imaginal discs and 96h AEL wing imaginal discs. **(A)** *D. melanogaster* and *D.*

mauritiana. In all samples, most genes have divergent expression due to variation in *trans*. **(B)** *D. simulans* and *D. mauritiana*. In eye-antennal imaginal discs, variation in *trans* is causing the differences in expression of most genes; in wing discs, more genes have compensatory regulation and are not significantly differentially expressed in the parental species.

I then wanted to know whether the genes with divergent expression are the same in the two studied tissues. For that I compared the genes that were found to be different due to each type of regulatory changes in eye-antennal imaginal discs and in wing imaginal discs at 96h AEL (Figure 4.3.9). The highest overlap was observed in the genes with expression changes due to variation only in *cis*, but the overlap was also rather high in the genes with compensatory regulation. A similar total number of genes had differences due to *trans* variation in the two tissues, although since many more total genes have this type of regulatory variation, the percentage is lower. Interestingly, the genes presenting variation in *trans* were the only ones that had different direction in the expression differences, i.e. 64 genes were up-regulated in *D. melanogaster* in eye tissues but up-regulated in *D. mauritiana* in wing tissue, and 134 genes were up-regulated in *D. mauritiana* in the eye-antennal imaginal disc but up-regulated in *D. melanogaster* in the wing disc. In the genes with divergent expression due to *cis* regulation, only one gene had higher expression in *D. mauritiana* in the eye but higher expression in *D. melanogaster* in the wing. This was the case in only two genes with compensatory regulation. From the few genes with *cis* x *trans* regulation, only one was commonly high in the two tissues in *D. mauritiana*.

This analysis in the *D. mauritiana* x *D. simulans* data gave similar results (data not shown). The direction of expression change in the two tissues happens only in the genes that present divergent gene expression due to changes in *trans* regulation.

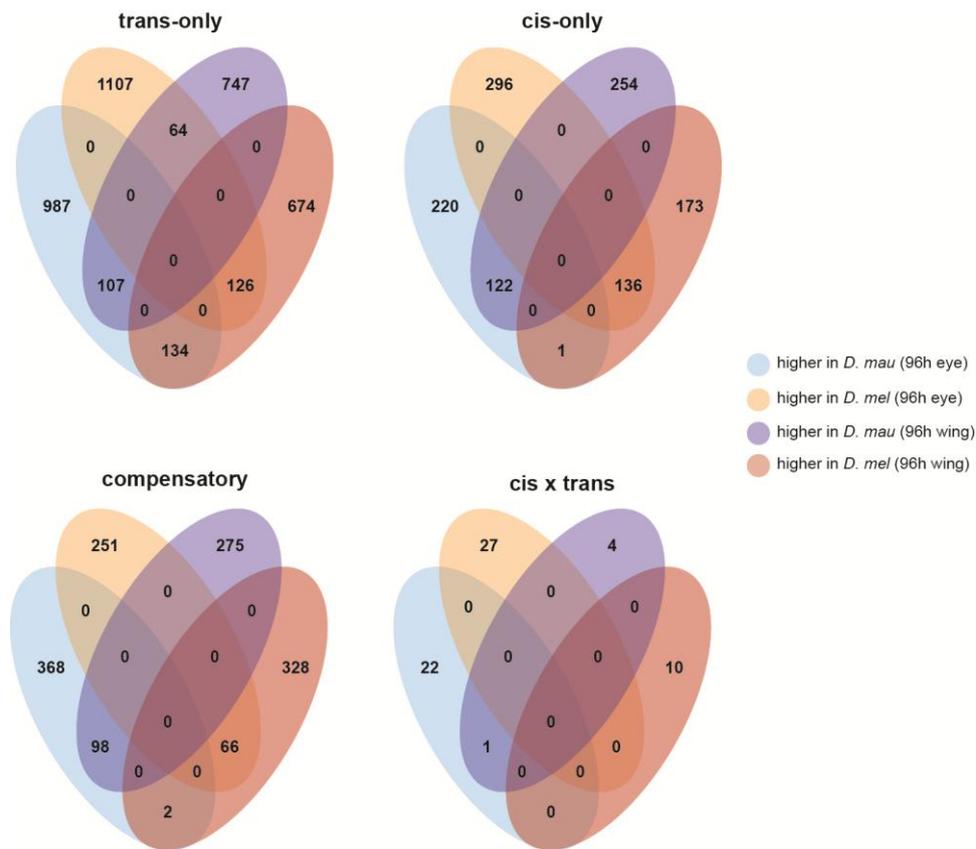


Figure 4.3.9. Overlap of regulation types between eye and wing tissue in *D. melanogaster* x *D. mauritiana* hybrids (96h AEL). The venn diagrams show the number of genes that are shared for each type of regulation, tissue and species. The background color indicates the tissue (light blue and light red show the number of genes in eye-antennal imaginal disc tissue and the dark blue and dark red show the genes in wing imaginal disc tissue) and the species with higher expression (shades of blue for *D. mauritiana* and shades of red for *D. melanogaster*). Only in genes with regulation in *trans* does the direction of the expression differences change in the two tissues.

Finally, I wanted to know what genes have divergent expression between *D. mauritiana* and *D. melanogaster* in the different tissues and investigated whether they had features in common. Thus I searched for enriched GO terms and upstream regulatory factors (Table 4.3.3). In the eye-antennal imaginal discs, the genes with higher expression in *D. melanogaster* due to regulation in *trans* are enriched for metabolic processes, while the genes with higher expression in *D. mauritiana* are involved in biological regulation and differentiation. I found very high enrichment of binding motifs for the transcription factor Pannier (NES = 9.48) in the genes that have higher expression in *D. melanogaster* due to variation in *trans* in the eye tissue at 96h AEL. The same factor is likely to regulate genes which have higher expression in *D. mauritiana* in the wing disc and whose higher expression is due to variation in *trans*. The genes with higher expression in *D. mauritiana* in the eye-antennal imaginal discs and with variation regulated in *trans* show enrichment for the binding motif of Ecdysone Receptor (NES = 6.92). Also the genes with higher expression in *D. mauritiana* in the eye

but with lower expression in the wing have enrichment for this upstream factor. Although there are less genes with divergent expression due to changes in *cis* regulatory regions and it is less informative to identify common upstream factors for these genes, significant enrichment for Pannier (NES = 7.57) was also present in the genes with higher expression in *D. melanogaster* in the eye.

Table 4.3.3. GO terms and transcription factor (TF) enrichment of *cis* and *trans* genes between *D. mauritiana* and *D. melanogaster* (96h AEL).

		# genes	GO terms*	TFs*
<i>trans</i> only	only eye higher <i>D. mel</i>	1,107	single-organism metabolic process, mitochondrial organization, nitrogen compound metabolic process	pnr (9.48)
	only eye higher <i>D. mau</i>	987	biological regulation, response to stimulus, neuron differentiation	EcR (6.92)
	only wing higher <i>D. mel</i>	674	localization, anion transport	-
	only wing higher <i>D. mau</i>	747	organonitrogen compound metabolic process, mitotic spindle elongation, centrosome duplication	pnr (5.57)
	both tissues higher <i>D. mel</i>	126	unannotated	Pmad (4.28)
	both tissues higher <i>D. mau</i>	107	-	ftz (5.20)
	eye higher <i>D.</i> <i>mel</i> , wing higher <i>D. mau</i>	64	oxidation-reduction process, chitin metabolic process, amino sugar metabolic process	bin (4.66)
	eye higher <i>D.</i> <i>mau</i> , wing higher <i>D. mel</i>	134	multicellular organismal development, cell morphogenesis, nervous system development	EcR (4.58)
<i>cis</i> only	only eye higher <i>D. mel</i>	296	-	pnr (7.57)
	only eye higher <i>D. mau</i>	220	biological regulation, metal ion homeostasis, macromolecule localization	caudal (3.69)
	only wing higher <i>D. mel</i>	173	biological regulation, response to stimulus	mef2 (4.83)
	only wing higher <i>D. mau</i>	254	mitotic spindle elongation	pnr (5.10)
	both tissues higher <i>D. mel</i>	136	UDP-glucose metabolic process	pnr (4.37)
	both tissues higher <i>D. mau</i>	122	metabolic process	kni (4.75)

*first three non redundant enriched GO terms

**first enriched TF (i-cisTarget).

4.4 Eye size variation in two closely related *Drosophila* species

Based on quantitative genetics approaches (Arif et al., 2013, unpublished data) 81 candidate genes located in a 1.1 Kb region on the X chromosome have been identified that could be responsible for the differences in ommatidia size between *D. simulans* YVF and *D. mauritiana* TAM16. Here I describe the work that I and Dr. Isabel Almudi (Oxford Brookes University, Oxford, UK) have carried out to combine next generation sequencing to detect differentially expressed genes and molecular and functional genetics to reduce this list of candidate genes.

4.4.1 Genes differentially expressed between species

I performed RNA-sequencing of *D. mauritiana* TAM16 and *D. simulans* YVF eye-antennal imaginal discs at 120h AEL (late LIII) (Table 3.2). I did not analyze the earlier time points because at this stage the retinal part is still similar between these two species (Arif et al., 2013). Therefore, the molecular differences that will give rise to the size differences have to occur at this time point or later.

The reciprocal re-annotation of the genomes of *D. simulans* and *D. mauritiana* (Torres-Oliva et al. in revision) allowed the comparison of gene expression levels of 13,239 genes between these species. First, I filtered out the genes that had very low expression in the two species, since these genes have been shown to mainly represent noise and disturb the overall analysis of differential gene expression (Anders and Huber, 2010). 9,144 genes had more than 1 read per million reads in at least 3 samples. Since these are very closely related species it could be that, for some genes, the expression differences are not very large or significant. Therefore, to reduce the chance of false positive genes, I applied two different methods to call differentially expressed genes, namely DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010) (Figure 4.4.1). According to DESeq 1,051 genes have significantly higher expression in *D. mauritiana* TAM16 and 1,030 have higher expression in *D. simulans* YVF ($p\text{-adj} < 0.05$). edgeR reports less significantly differentially expressed genes, 773 higher in *D. mauritiana* TAM16 and 678 higher in *D. simulans* YVF (FDR < 0.05). These results and the MA plot in Figure 4.4.1 show that edgeR is a more conservative approach.

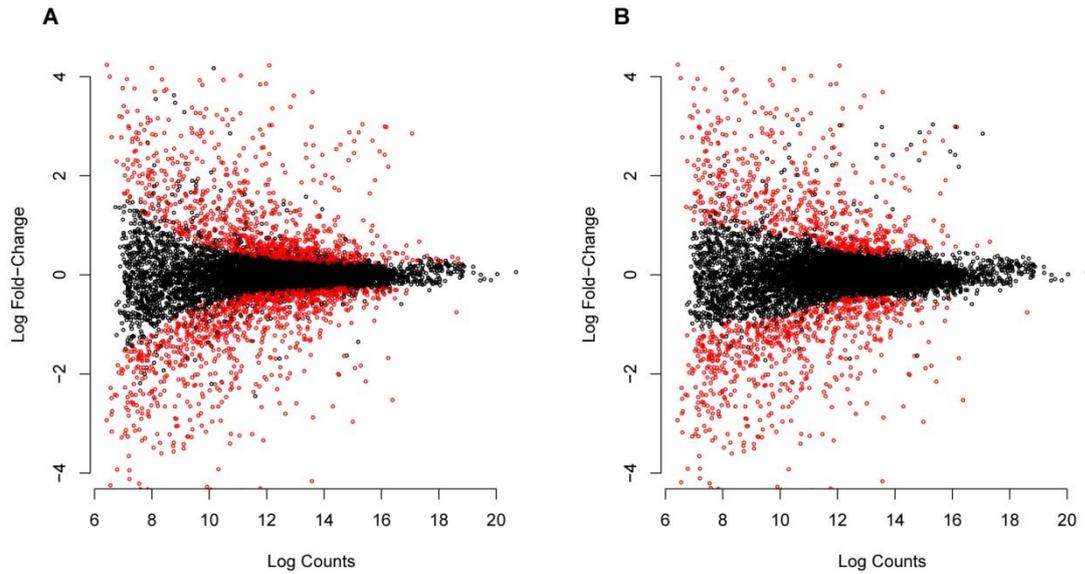


Figure 4.4.1. MA plot of differential gene expression analysis. Plot of expression ratios (y-axis) vs. mean of the average count (x-axis). Each point represents one gene, red dots are significantly differentially expressed genes. Genes with positive log fold-change have higher expression in *D. simulans* YVF compared to *D. mauritiana* TAM16 and viceversa. **(A)** DESeq results ($p\text{-adj} < 0.05$). **(B)** edgeR results ($FDR < 0.05$).

We then combined this differential expression data with the QTL mapping information. The QTL for eye size between *D. mauritiana* TAM16 and *D. simulans* YVF has been mapped to a region between 7.4 Kb and 8.5 Kb in the chromosome X of *D. simulans* (Figure 4.4.2A; FlyBase assembly (Hu et al., 2013), Scf_X). Of the 81 genes in this region (Figure 4.4.2B), 76 are expressed (more than 1 read per million reads in at least 3 samples) and only 14 genes are differentially expressed (Table 4.4.1) according to at least one of the used methods. DESeq calls all these 14 genes significantly differentially expressed. As it was already shown before, edgeR is more conservative and only calls as differentially expressed 8 of these 14 genes and no additional one (Supplementary Table 5). The following analyses were performed on these 14 candidate genes.

in the anterior part of the eye field, also with equal expression in the two species. In *D. simulans* *Sptr* shows an expression domain in the dorsal side of the face region and a smaller domain at the ventral region adjacent to the morphogenetic furrow; in *D. mauritiana* no staining could be observed. *Es2* presents ubiquitous expression in the disc in *D. simulans*, while in *D. mauritiana* it does not seem to be expressed in the most posterior region of the eye disc. *oc* is expressed in a clear domain where the ocelli develop (dorsal side of the face region) and in the posterior region of the eye disc; in *D. mauritiana* this region is slightly wider at this stage (Supplementary Figure 12).

In addition, we also performed a functional analysis of the 14 candidate genes in the model species *D. melanogaster*. Using the UAS/Gal4 system in combination with Dicer expression (Dietzl et al., 2007) we knocked-down the different candidates using an eye-specific driver (GMR) and scored adult eye morphology (Table 4.4.1). GMR corresponds to a response element from the gene Rh1 opsin, which drives expression in all cells posterior to the morphogenetic furrow (Freeman, 1996). Crossing the flies at 28°C, most of the candidates gave rise to very weak or no phenotype (*Tbh*, *CG1632*, *Galc*, *Es2*, *Sptr*, *sni*, *CG1885* and *CG2004*; phenotype only visible under electron microscope, few irregular ommatidia detectable) or they produced no offspring due to an unsuccessful cross (*CG10958*, *CG1575* and *Ppt1*; since these genes are not expressed in the posterior region of the eye disc, the crosses were not repeated). Only *CG2254* and *spirit* gave mild phenotypes such as slightly rough eye. *Oc* knock-down clearly resulted in the strongest rough eye phenotype (Supplementary Figure 13).

We also performed the crosses at 25°C. In that case, most of the studied candidate UAS-RNAi lines did not have a phenotype, only *sni* and *CG1632* (weak), one of the lines of *spirit* (mild) and two of the lines of *oc* and one of the lines of *Es2* (severe) (not shown).

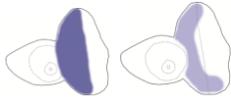
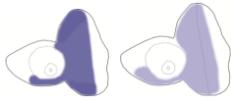
In accordance with the *in situ* stainings, the lines with strongest phenotypes (*spirit*, *Es2* and *oc*) are the genes that appear to be expressed in the posterior region of the retinal field: *spirit* and *Es2* show ubiquitous expression and *oc* is expressed in the more posterior part of the eye field (Table 4.4.1, Supplementary Figure 12).

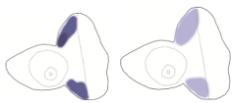
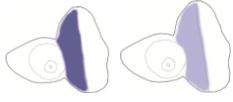
To note is that the strongest phenotype at 28°C was obtained with the control flies that only contained the driver construct GMR-Gal4. Most likely, due to the fact of overexpressing Gal4 in the absence of a promoter where it can bind, which has been shown to produce unspecific phenotypes before (Cao et al., 2008). However, it sheds a question to the experiment, which should probably be repeated with another driver

construct. But the control GMR>>GFP did not produce a phenotype. When performing the crosses at 25°C no phenotype was observed in the GMR>>GFP control and it was weaker in GMR-Gal4.

In summary, we identified *oc* as a candidate that is expressed in the developing photoreceptors at late LIII stages and for which RNAi resulted in the most consistent relatively strong compound eye phenotype. This is also the only candidate gene with known roles in eye development, especially in ocelli development (Royet and Finkelstein, 1995), photoreceptor subtype differentiation by the regulation of *rhodopsin* expression (Tahayato et al., 2003) and to be involved in photoreceptor maturation (Fichelson et al., 2012).

Table 4.4.1. Summary of candidate genes.

Gene ID	Gene Name	DESeq Log2FC	DESeq p-adj	Expression pattern**	RNAi †	GO Terms (Biological Process)
FBgn0030004*	CG10958	-1.02E+00	1.33E-14		n.a.	-
FBgn0029994	CG2254	-2.57E+00	4.40E-08		++	oxidation reduction
FBgn0010329	Tbh	-2.63E+00	4.89E-07		+	regulation of neurotransmitter levels, histidine metabolic process, cell-cell signaling, gamete generation, memory, mating, response to ethanol
FBgn0030051	spirit	-7.47E-01	9.77E-06		++	proteolysis, defense response, immune response, regulation of Toll signaling pathway, positive regulation of cell communication
FBgn0030027*	CG1632	-6.33E-01	2.11E-05		+	proteolysis
FBgn0040319	Gclc	-1.45E+00	8.63E-05		+	peptide metabolic process, sulfur metabolic process, cellular response to DNA damage stimulus, cofactor biosynthetic process
FBgn0023506	Es2	5.37E-01	2.99E-04		+	-
FBgn0014032*	Sptr	4.85E-01	5.76E-04		-	tetrahydrobiopterin biosynthetic process, nitrogen, oxidation reduction

FBgn0029999	CG1575	3.81E-01	1.21E-02		n.a.	-
FBgn0030026*	sni	7.40E-01	1.30E-02		+	oxidation reduction
FBgn0030066*	CG1885	4.00E-01	2.61E-02		-	heterocycle biosynthetic process, tetrapyrrole biosynthetic process, nitrogen compound biosynthetic process
FBgn0004102	oc	-4.23E-01	3.41E-02		+++	compound eye photoreceptor cell diff., regulation of transcription, zygotic determination of A/P axis, metamorphosis, adult walking behavior, ocellus devel., neuron diff., brain segmentation, rhabdomere devel., cell fate commitment, regulation of RNA metabolic process
FBgn0030060	CG2004	2.92E-01	3.50E-02		+	-
FBgn0030057	Ppt1	-3.25E-01	3.85E-02		n.a.	protein depalmitoylation, aging, determination of adult life span, lipoprotein metabolic process

* I have performed the analysis of these genes. The other genes have been analysed by Dr. Isabel Almudi (Oxford Brookes University, Oxford, UK).

** Expression pattern in *D. simulans* in darker shade (left) and, when available, *D. mauritiana* in lighter shade (right).

† “-“: no phenotype; “+“: weak phenotype; “++“: mild phenotype; “+++“: severe phenotype, “n.a.“: unsuccessful cross.

4.4.3 Coding sequence divergence

Changes in gene expression regulation are thought to be more likely the reason for the evolution of morphological differences, especially between closely related species (Carroll, 2008). However, we cannot discard that the observed differences in ommatidia size are caused by differences in the coding sequence of genes. Therefore, I performed pair-wise alignment of each pair of orthologs for the 76 genes present in the QTL region and that have some gene expression as measured by RNA-seq (Supplementary Table 5). 13 orthologs have 100% identity and only 2 orthologs (fs(1)h and CG10555) have less than 90% identity in their coding sequences. However, these two genes have high percentage of repetitive sequence, such as very long glutamine stretches, which are known to cause sequencing and assembly problems. Two genes have more than 80 aminoacid changes between the two species (Trf2 has 82 single nucleotide polymorphisms and Nrg has 84), although due to the fact that they have long sequences they have high identity percentage (Trf2 90.4% and Nrg 93.6%).

4.4.4 Optical sections of *Drosophila* heads

Although we know that the eyes of these two species differ due to differences in ommatidia size, this was measured by calculating the area of the lens of five central ommatidia on the outer surface of the eye (Posnien et al., 2012). This means that we still have no information about the underlying nature of this morphological difference at the cellular level. It could be due to longer or shorter ommatidia cells, or because these are wider or narrower in one of the species. We also do not know if the cells contributing to these differences are the photoreceptors, the pigment cells or the supporting cells that secrete the ommatidia lens (Waddington and Perry, 1960). To better understand all of this, my aim was to image the interior of the adult eyes and measure these features.

Using a recently published protocol (Smolla et al., 2014) I cleared the heads of adult flies of both *D. simulans* YVF and *D. mauritiana* TAM16 and scanned them using the laser scanning microscope taking advantage of the auto fluorescence of the cuticle (Figure 4.4.3A). With this method I could perform precise measurements of different features of the eye (Figure 4.4.3B). The number of ommatidia in the central row (from dorsal to ventral) is significantly higher in *D. simulans* ($p=0.0222$) (Figure 4.4.3C) ($n=6$ for *D. mauritiana* TAM16 and $n=7$ for *D. simulans* YVF in all measurements). Since in some heads the lenses were no longer attached to the ommatidial clusters, I measured the ommatidia length from both the top of the ommatidia and from the base of the lens (red and green lines in Figure 4.4.3B,

respectively). Both measurements gave very similar results, and showed that the ommatidia length is not significantly different between the two species ($p=0.210$ for ommatidia length and $p=0.110$ for length from the lens), although the values in *D. mauritiana* TAM16 had a higher mean (Figure 4.4.3D and E). It was also possible to measure the eye diagonal, which is the distance between the most dorsal and the most ventral margins of the eye (blue line in Figure 4.4.3B). In this case, *D. mauritiana* TAM16 had a higher mean for this value, but the difference is not significant ($p=0.187$) (Figure 4.4.3F). Finally, I measured the ommatidia width, both as the width of the lens and as the distance between the adjacent pigment cells (orange and yellow lines in Figure 4.4.3B, respectively). Both measurements are significantly larger in *D. mauritiana* TAM16 ($p=0.0051$ for lens width and $p=0.0121$ for pseudocone width) (Figure 4.4.3G and H), coinciding with the results that this species has larger ommatidia lens surface (Posnien et al., 2012).

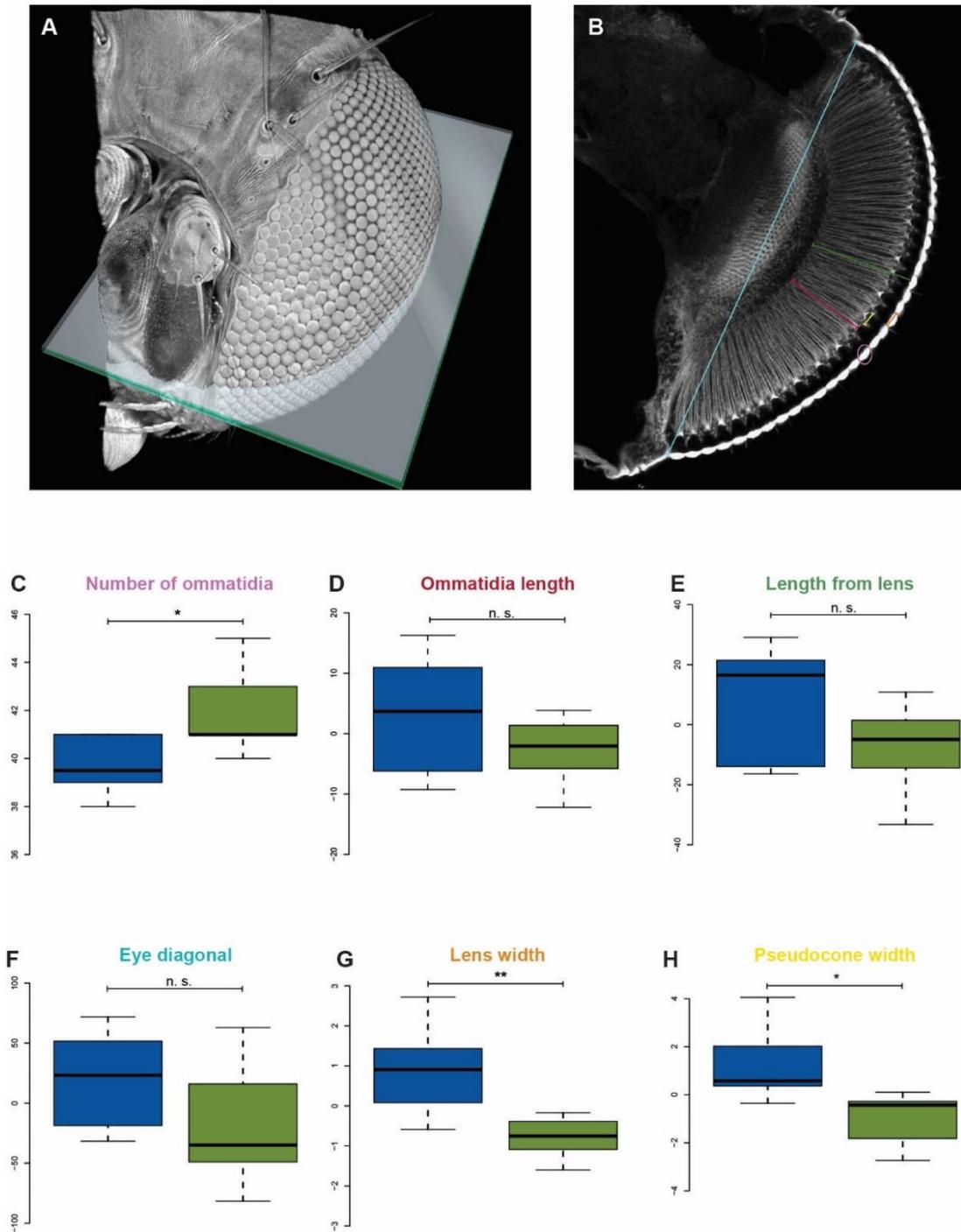


Figure 4.4.3. Head optical sections. (A) Reconstruction of a *D. simulans* head from 50 optical sections. The crossing surface indicates the location of the section used for the measurements. (B) Example of a section used to measure eye and ommatidia structures. For all analyzed eyes, the first section where the brain lamina was visible (from the dorsal side) was used. Colored lines and circles indicate the landmarks that were used for the measurements. Colors correspond to the title of the plots in (C-H). (C-H) Boxplots of the measurements indicated in (B). Values are normalized for body size using the reciprocals of the correlation with tibia length. Blue boxes correspond to *D. mauritiana* TAM16 measurements and green to *D. simulans* YVF. “n.s.”: not significant; “*”: p-value < 0.05; “**”: p-value < 0.005.

5 Discussion

5.1 New regulatory interactions governing *Drosophila* head development

The developmental transcriptomics analysis of eye-antennal imaginal discs of *Drosophila* has provided new insights into the gene expression changes that define the development of this tissue. Here I will discuss the identification of biological processes and different dynamic expression profiles and how this has been used to discern previously known and new transcription factors involved in head and visual system development. In particular, the finding of a new role of the transcription factor Hunchback in retinal glia cells development is extensively discussed in accordance with the experimental results obtained.

5.1.1 **Dynamic gene co-expression describes eye-antennal imaginal disc developmental events**

The pair-wise differential gene expression between developmental stages clearly shows that the most pronounced transition in the eye-antennal imaginal disc happens as larvae progress from LII stage into LIII stage. Between these two time points, 50% of the expressed genes show significant differential expression. At late LII stage the cells in the eye-antennal imaginal discs are mostly involved in metabolic processes and generation of energy (Figure 4.1.2A). This can be explained by the fact that these cells are mainly in a proliferative state, as the discs have to grow to immensely increase their size (Kenyon et al., 2003; Kumar and Moses, 2001). Actually, it is this growth what allows the posterior and anterior morphogen gradients of Wingless (Wg) and Decapentaplegic (Dpp) to separate and initiate the compartmentalized expression of *eyes absent* (*eya*) and *sine oculis* (*so*) in the posterior margin of the disc, which will trigger the events of retinal differentiation (Domínguez and Casares, 2005; Kenyon et al., 2003). Dpp is necessary for the activation of early retinal genes (Curtiss and Mlodzik, 2000; Kenyon et al., 2003), while Wg represses this expression (Hazelett et al., 1998). Retinal differentiation can only start when the disc has grown enough to create a Wg-free region on the posterior margin (Domínguez and Casares, 2005; Kenyon et al., 2003). Correspondingly, at 96h AEL genes related to cell differentiation, nervous system development, pattern specification and also compound eye development are significantly up-regulated. An interesting observation is that many GO terms related to eye development can be found with high enrichment score, while no GO

term specific for antenna, maxillary palps or head cuticle were found. This shows that the research on eye specific development has been much more extensive than that on the other organs that develop from the same imaginal disc. The transition from mid LIII stage (96h AEL) to late LIII (120h AEL) is less pronounced, although up to 22% of genes shift their expression levels. Interestingly, in this transition again genes related to metabolism and energy production are down-regulated (Figure 4.1.2B). This can be explained by the fact that at 96h AEL the disc has not yet reached its final size, and cells anterior to the morphogenetic furrow still proliferate. Also directly behind the morphogenetic furrow one last synchronous cell division takes place to give rise to the last cells of the photoreceptor clusters (R1, R6 and R7) (Baonza et al., 2002). The GO terms of the up-regulated genes are also similar to those enriched in the genes up-regulated in the first transition, but in this case some terms related to later processes are listed, such as R7 cell differentiation or pigment metabolic process. Also genes related to leg disc pattern formation are enriched, which can be explained by the fact that the pathways involved in leg and antenna development are very similar (Abu-Shaar and Mann, 1998; Campbell and Tomlinson, 1998; Dey et al., 2009).

An even better resolution of the different processes taking place during eye-antennal imaginal disc development can be obtained by the co-expression gene clustering using HTScluster (Rau et al., 2015) (Figure 4.1.3). For example, cluster 7 groups genes that are similarly highly expressed at 72h and 96h AEL, and their expression decreases at 120h AEL. The known genes in this cluster have been described to be related to DNA replication and cell cycle (Table 4.1.2), which corresponds with the fact that active proliferation is taking place at these stages (Baonza et al., 2002). Thus, other genes that have been grouped in this cluster but for which no previous knowledge is available are likely also related to these biological functions. Another interesting cluster is cluster 5, which groups genes with higher expression at 72h and 120h AEL, but down-regulated at mid LIII stage (Figure 4.1.3). These are only 283 genes and they do not share enriched GO terms, but they could be involved in processes related to molting and preparation for stage transitions.

With the co-expression clustering, also the early expressed genes (clusters 1, 2, 6 and 8) are divided more precisely according to how pronounced their changes in expression are (Figure 4.1.3). Interestingly, although these clusters contain more than 2,500 genes altogether, very few genetic interactions are known among the genes of each cluster (Table 4.1.2), and this provides a niche for new connections and key regulators to be found. A

better resolution of the events and interactions taking place at this early stage could be obtained by sequencing the transcriptome specific for the different imaginal discs' regions. This could be done by independent driver lines followed by fluorescence activated cell sorting (FACS (Hewitt et al., 2006)) and RNA-seq to obtain the antennal region transcriptome (e.g. using *cut*-Gal4), the eye region transcriptome (e.g. using *ey*-Gal4) and the complete disc (e.g. using *bth*-Gal4) to reduce noise and make sure that only genes expressed in the eye-antennal imaginal disc are sequenced and not those present in surrounding tissue.

Similarly, genes up-regulated in the later stages are separated in more specific clusters, and most of the enriched GO terms are related to differentiation and neuron and eye development. Cluster 9 contains genes with similarly high expression at 96h AEL and 120h AEL (Figure 4.1.3). This cluster contains the most known genetic interactions among its members, and it includes the well-known developmental pathways EGFR, Notch and cell cycle related genes (*CycE* hub) (Figure 4.1.4C). Cluster 11 contains genes which are steadily up-regulated. Among them, *Delta* (*Dl*) (Figure 4.1.4B), which is one of the Notch receptor ligands (Baker, 2000) and has different roles in eye development (Frankfort and Mardon, 2002; Kumar and Moses, 2000), is found as one of the hub genes. Also *anterior open* (*aop*) (also known as *yan*), which is described to repress photoreceptor differentiation (O'Neill et al., 1994) and also to determine R3 photoreceptor type (Weber et al., 2008) is present in this cluster. Cluster 4 groups together genes that are highly expressed only at late LIII stage, and correspondingly shows enrichment for genes involved in pigmentation and pupariation (Table 4.1.2).

These dynamic developmental expression profiles are very similar when adding the closely related species *D. mauritiana* and *D. simulans* to the analysis (Figure 4.3.2), and it shows that these processes are greatly conserved during eye-antennal imaginal disc development. 13 co-expression clusters are predicted when using only *D. melanogaster* data and also when using data from the three different species. Only one more cluster with high expression at 72h AEL and low expression in the later stages is predicted when using the three species (cluster 8, Figure 4.3.2). In contrast, the cluster with higher expression at 72h AEL and 120h AEL and slightly lower at 96h AEL (cluster 5, Figure 4.1.3) is only predicted when using only *D. melanogaster* data. All other clusters predicted with the three species transcriptomic data reproduce exactly the profiles obtained for *D. melanogaster* only. The GO terms are again very specific for each process and can be followed in time when the clusters are ordered by the stage of gene expression. Some of the highly enriched GO terms in this case are more specific than when using only *D. melanogaster* and define the

underlying subnetworks with higher resolution. Thus the clustering of expression profiles recapitulates remarkably well the different events taking place during the studied stages during eye-antennal imaginal disc development, which are highly conserved across the studied closely related species.

5.1.2 Enriched *cis*-regulatory elements in co-expressed genes identify upstream transcription factors

Clusters of co-expressed genes can also unravel co-regulatory upstream factors. I have used the method i-cisTarget (Herrmann et al., 2012) to identify common *cis*-regulatory elements in each of these gene clusters. This method can perform this search and later combine it with available ChIP-seq datasets to find specific experiments where a transcription factor was found to significantly bind a large number of the genes in that cluster. For instance, the results of this analysis using only *D. melanogaster* indicate that Ecdysone receptor is enriched to regulate a significant number of genes in clusters 5, 11 and 4 (Table 4.1.3), corresponding with the ecdysone hormone pulses before larval molting and pupariation (Li and Bender, 2000). These results could indicate that this hormone is also activating signaling cascade in the eye-antennal imaginal disc at the onset of these transitions.

Interestingly, when using the three *Drosophila* species, the genes of the clusters with early expressed genes are more enriched for *cis*-regulatory elements of the transcriptions factors studied by the modENCODE project (Celniker et al., 2009) such as Nejire and Caudal (Figure 4.3.2). Nejire is a co-factor known to be involved in many processes of eye development and patterning (Kumar, 2004). This zinc-finger DNA binding protein is a co-activator that can act as bridge for other transcription factors to bind specific enhancer elements (Dai et al., 1996; Kwok et al., 1994; McManus and Hendzel, 2001). This can explain why I find it to regulate such a large number of target genes. It could be that Caudal also plays a similar role in this process. It has been indeed described that Caudal is a downstream core promoter activator (Juven-Gershon et al., 2008) and very recently it has been found that it works together with Nejire to promote the expression of the homeobox gene *fushi tarazu* (*ftz*) (Shir-Shapira et al., 2015). My results suggest that they could also be acting together during *Drosophila* eye-antennal imaginal disc development, and *ftz* is also found enriched to regulate genes in a cluster of early expressed genes (cluster 1, Table 4.1.3).

Most of the other transcription factors identified in clusters with up-regulated genes have been already described as regulators of different processes of eye development. For

instance, a significant number of Sloppy-paired 1 target genes are up-regulated at LIII stage (cluster 11, Table 4.1.3) and this transcription factor is known to play a critical role in establishing dorso-ventral patterning of the eye imaginal disc (Sato and Tomlinson, 2007). A function of Daughterless (identified in cluster 4, Table 4.1.3) is also described; it is expressed in the morphogenetic furrow, it interacts with Atonal and is necessary for proper photoreceptor differentiation (Brown et al., 1996). Snail (cluster 1 and 4, Table 4.1.3) and Twist (cluster 11, Table 4.1.3) were identified in a screen for retinal determination genes as possible repressors of *dachshund* expression (Anderson et al., 2006) and my results could indicate that they regulate also other genes during eye-antennal imaginal disc development.

Another remarkable finding is that when using the three closely related species, three consecutive clusters (subcluster 3, 4 and 7, Figure 4.3.2) include genes enriched to be regulated by Pannier, Dorsocross2 and Pmad, all of which were studied by the Furlong Lab (Junion et al., 2012) and chosen because of their involvement in cardiac cell fate specification. This could indicate that the genetic networks (including transcription factors and many of their target genes) involved in early mesoderm specification are later also necessary to regulate cell cycle processes, patterning and development of eye-antennal imaginal disc tissue. Both processes are known to require Wg and Dpp signaling (Lee and Frasch, 2005; Royet and Finkelstein, 1996). My data would indicate that a large number of the underlying target genes of both processes could also be shared.

These are very promising results, as they show that, although the ChIP-seq experiments that identified the direct interaction of these transcription factors with their target genes were mostly performed during embryo stages, they can be used to identify upstream regulators in a completely different tissue. Clusters 1 and 2 (Table 4.1.3) retrieve the transcription factor Caudal from a ChIP-seq experiment performed in adult flies (Celniker et al., 2009) but do not identify this transcription factor from an experiment performed in embryos (Li et al., 2008). This could indicate that Caudal has very different downstream targets during embryogenesis from its target genes at later stages. However, it may also indicate that the parameters and thresholds used in the ChIP-seq experiments performed by these two groups are very different. This could also be the case of the transcription factor Pannier, which in cluster 12 has more than 1,200 highly ranked regions. Therefore, efforts like the modENCODE project (Celniker et al., 2009) are of vital importance in order to standardize the protocols and criteria to perform these experiments.

The correct identification of transcription factors known to be implicated in head and eye development in my dataset supported the analysis pipeline and the strategy to identify upstream orchestrators of these processes. Therefore, it became interesting to investigate the finding that the Myocyte enhancer factor 2 (Mef2) and Hunchback (Hb) could be also involved in this process. These two transcription factors have well described roles in other developmental processes, but so far no function in the development of eye-antennal imaginal discs had been reported.

5.1.2.1 A potential role of Mef2 in eye-antennal imaginal disc development

The MADS-box transcription factor Mef2 is crucial for the development of heart and muscle tissues (Gunthorpe et al., 1999). It is expressed in all mesodermal cells during blastoderm stages and its expression gets restricted by the action of the transcription factors Twist and Tinman (Lilly et al., 1994; Nguyen et al., 1994). I have identified many Mef2 target genes up- and down-regulated during eye-antennal imaginal disc development (clusters 2, 8 and 11). Using available Gal4 driver lines I could show that enhancer regions near its locus drive expression in some cells at the most anterior end of the discs (Supplementary Figure 2). Although this region is not considered part of the disc proper, but rather belongs to the peripodial membrane, this region was also dissected together with the discs that were sequenced and it could belong to future head muscle cells. Although no clear signal could be observed using α -Mef2 antibodies (not shown), some recent findings could hint towards an important role of this transcription factor in eye development. It has been recently reported that Mef2 is implicated in circadian behavior, as it is necessary for the proper fasciculation-defasciculation cycle of neurons (Sivachenko et al., 2013) through one of its target genes *fasciclin 2* (*fas2*), which is expressed in some photoreceptor neurons (Figure 4.1.16A and Mao and Freeman, 2009). Additionally, a recent transcriptomics study of larval eye and adult ocelli found that *mef2* is expressed in the photoreceptors of both eye types, although the authors did not investigate this finding further (Mishra et al., 2016). These findings certainly encourage additional research on the possible role of Mef2 in photoreceptor cell development.

5.1.3 Description of a new role of Hb in retinal glia development

The comprehensive analysis of developmental high-throughput gene expression data in combination with the identification of key upstream regulators also suggests that Hb may play an important role during eye-antennal imaginal disc development. Hb is a C2H2 zinc-

finger transcription factor that has been largely studied in *Drosophila* (Tautz et al., 1987). It was first identified as a gap segmentation gene due to its role in the very early steps of anterior/posterior axis determination, where it is regulated by the maternally expressed gene *bicoid*, which specifies anterior fate (Lehmann and Nüsslein-Volhard, 1987; Nüsslein-Volhard and Wieschaus, 1980). Later it was also found that Hb regulates temporal neuroblast identity during embryogenesis, as it determines first-born identity in neural lineage (Grosskortenhaus et al., 2005; Isshiki et al., 2001). Here I have revealed a new role of this transcription factor in the development of a subtype of retinal glia cells.

Using immunostaining and reporter gene expression I confirmed that *hb* is indeed expressed in two large cells in the posterior margin of the eye-antennal imaginal discs (Figure 4.1.5). Further co-expression analysis with glia cell markers indicated that these cells are retinal sub-perineural glia cells known as carpet cells (Silies et al., 2007) (Figure 4.1.8).

There are only two carpet cells in each eye imaginal disc and these cells have quite unique features. Like other sub-perineural glia cells they have very large, polyploid nuclei and huge cell bodies. The carpet cells work as a scaffold of other retinal perineural glia, which are still undifferentiated and migrate to find the nascent axons of the differentiating photoreceptors (Silies et al., 2007). When perineural glia cells contact these axons, they differentiate into wrapping glia cells and then they enwrap the axons to participate in their projection to the brain lobes (Hummel et al., 2002). It is also thought that carpet cells are necessary to prevent the over migration of perineural glia cells anteriorly from the morphogenetic furrow (Silies et al., 2007). Importantly, carpet cells are the only sub-perineural glia cells that migrate through the optic stalk into the eye-antennal imaginal discs during LII stage. They express the G-protein coupled receptor Moody (Bainton et al., 2005) and form septate junctions with other surface glia, which contribute to the establishment of the blood-brain barrier (Schwabe et al., 2005). To study the possible role of Hb in these cells I have tested these features in the cells that express *hb* and also what happens to carpet cells when the expression of *hb* is affected.

5.1.3.1 *hb* expression is necessary for the presence of polyploid carpet cells in the eye disc

The loss of *hb* expression in the carpet cells, both by the use of RNA interference and by a temperature sensitive null mutant, has reduced the presence of the characteristic large nuclei of these glia cells in the eye discs (Figure 4.1.11). A stronger RNAi effect can be observed when a *moody*-Gal4 driver is used in contrast to *repo*-Gal4. This is probably due to

the fact that Moody is a G protein-coupled receptor that is constantly required in the subperineural glia cells to form septate junctions (Bainton et al., 2005) and therefore it is highly expressed in these cells. The effect of loss of Hb function was also slightly stronger when *hb^{ts}* mutant flies were transferred to the restrictive temperature during LI larva stage rather than at later stages, indicating that *hb* is expressed in carpet cells already at the first larval stage. Since Hb is necessary for normal embryonic development (Lehmann and Nüsslein-Volhard, 1987), a potential role of this transcription factor in carpet cells at embryonic stages could not be studied because no larvae hatched after *hb* was knocked-out during embryogenesis.

It is still not clear where carpet cells originate from. Although most publications indicate that these cells originate in the optic stalk, to affirm this they cite Choi and Benzer 1994. In this publication, the authors indeed observed the presence of carpet cells at late LII stage with an enhancer trap line (M1-126). However, they did not analyze earlier larvae. It is still not clear if carpet cells indeed originate in the optic stalk or if, alternatively, they originate from a pool of neuroblasts in the neuroectoderm during embryogenesis (reviewed in Homem and Knoblich, 2012) or in the optic lobes (reviewed in Apitz and Salecker, 2014). The fact that in loss of *hb* experiments we can observe in some cases only one polyploid cell nucleus and in some cases no polyploid cell nucleus could indicate that the two carpet cells originate independently from each other. The use of the newly analyzed driver lines VT038544 and VT038545, which drive expression only in the carpet cells glia subtype, can help to better understand the origin of these cells.

An additional phenotype observed in *hb* loss of function larvae is the lack of glia cells in small regions of the retinal field (Figure 4.1.12). This was accompanied by the presence of unorganized axon bundles that did not seem to properly project into the optic stalk. This was observed in eye discs in which carpet cell-like nuclei were not present. A possible explanation for the patches lacking glia cells could be the absence of carpet cell surface to work as support layer for perineural glia cells. It has been indeed described that in the absence of glia cells, projecting axons are not able to enter the optic stalk or get directed to it (Rangarajan et al., 1999). To be sure that areas of the retinal field are lacking perineural glia cells, precise glia cell quantification analyses should be performed.

I have used a Repo antibody to detect the presence of the large polyploid nucleus of carpet cells. Although the number of polyploid nuclei is drastically reduced upon loss of Hb function, I cannot rule out that the carpet cells are still there but have, for example, a smaller nucleus and cannot be distinguished from the other perineural glia cells.

Additionally, other studies have shown that carpet cell ablation or a reduction of their size causes over migration of other glia cells anterior to the morphogenetic furrow (Silies et al., 2007; Yuva-Aydemir et al., 2011). The fact that I do not observe this phenotype in the loss of Hb experiments could indicate that carpet cells are not completely missing. One possibility is that the migration abilities of carpet cells are reduced but their cell margins can still grow to work as boundary to prevent the migration of perineural glia cells anteriorly to the morphogenetic furrow. When performing all loss of function experiments, I separated the eye-antennal imaginal discs from the brain by cutting the optic stalk. I only scored the presence of carpet cell-like polyploid nuclei on the posterior edge of the eye disc, but I cannot rule out that these nuclei were present at the top of the optic stalk or even still close to the brain. In many cases, only one carpet cell could be observed in the eye disc, and this was often larger and located in the midline of the eye field. In these cases, also no perineural glia cell over migration could be observed, what might indicate that this single carpet cell was able to extend its cell margin to probably cover the complete retinal field. In future experiments, the use of a reporter line that marks the cell membrane of these cells will also help elucidating whether carpet cells are present in the eye disc or not when they do not express *hb*.

5.1.3.2 *hb* expression can induce carpet cell-like behavior in other glia cell types

hb misexpression experiments in different retinal glia subtypes have been rather useful. Although the results of driving ectopic *hb* expression in perineural cells are not conclusive due to the small number of individuals that could be analyzed, it seems like it might have induced a carpet cell-like behavior. The presence of many glia cells with large nuclei in the optic stalk could be indicative of this. This would also mean that the expression of *hb* under control of the perineural glia cell specific driver c527-Gal4 (Ito et al., 1995) is early enough to still induce carpet cell behavior in these cells. Perineural glia cells are still proliferating and undifferentiated (Rangarajan et al., 1999, 2001) and therefore expression of *hb* could still change their fate into ectopic carpet cells.

Wrapping glia are differentiated glia cells and therefore *hb* misexpression in these is probably too late to affect their fate. Accordingly, misexpression of *hb* in wrapping glia cells did not affect larvae survival and therefore the resulting phenotypes could be better analyzed and were more consistent (Figure 4.1.13). It is likely that the cell nuclei that can be observed between the axon bundles in the optic stalk belong to wrapping glia cells that misexpress *hb*. This would indicate that the expression of *hb* results in an over migration of

these cells, which normally remain in the eye disc and only their extended cell margins project to the brain lamina or medulla to accompany the photoreceptor axons (Hummel et al., 2002). Alternatively, the cell nuclei present inside the optic stalk could belong to the perineural glia cells that normally form a monolayer around the complete cluster of axonal projections (Hummel et al., 2002). This could be caused by an improper coverage of the individual axon bundles by wrapping glia cells that could produce a “leakage” of perineural glia cells. The use of perineural glia cell type specific cell markers could help clarify this phenotype in case these are the cells that are found between the axon bundles.

5.1.3.3 Hb expression in carpet cells is necessary for blood-eye barrier formation

Interestingly, the loss of *hb* expression in carpet cells also affected the integrity of the blood-eye barrier (Figure 4.1.14). This effect was not as striking as in previously published *moody* mutant flies. Yet this could be expected as *moody* mutations affect all sub-perineural glia cells, the carpet cells and those covering the brain (Bainton et al., 2005). It has been shown that during pupation, carpet cells migrate back into the optic stalk to the brain lobes, and by mid-pupa stages they are already located at the base of the brain lamina (Edwards et al., 2012). In the adult, they are also located there and, together with other sub-perineural glia cells, they form septate junctions that isolate the brain and retina from the hemolymph (Carlson et al., 2000). The experiments of blood-brain barrier integrity have only been performed using *moody* driven RNAi because *hb^{ts}* animals that grow at the restrictive temperature during larval stages do not survive to adulthood. The blood-brain barrier is already established by the end of embryogenesis, at least the layer formed by sub-perineural glia cells (Beckervordersandforth et al., 2008; von Hilchen et al., 2013). During larval stages only perineural glia cells continue to proliferate (Awasaki et al., 2008). However, sub-perineural glia cells can still undergo large migration and growth processes after larval hatching (Choi and Benzer, 1994). This means that in *hb^{ts}* animals, which I kept at 18°C during all embryogenesis, the sub-perineural glia cells are already present and their cell membranes probably also form septate junctions with adjacent sub-perineural glia cells. It would be informative to try to grow *hb^{ts}* larvae at the restrictive temperature only shortly enough for them to be able to develop into adults and repeat the blood-eye barrier assay with these individuals. This could help to elucidate if *hb* is needed only early during carpet cell development to preserve the structure of the blood-eye barrier or if it is also necessary later for proper migration of these cells into the base of the brain lamina (Edwards et al., 2012). Blood-brain barrier mechanisms are of foremost importance for all metazoan organisms due to its pivotal role in maintaining the correct physiological conditions in the

central nervous system. Also in vertebrates, glia cells and especially astrocyte glia are the main components of this barrier (Iadecola and Nedergaard, 2007). Thus the study of the function of sub-perineural glia cells in blood-brain barrier formation in the invertebrate model *D. melanogaster* can be of great interest to gain insight into central nervous system physiology and disease studies (DeSalvo et al., 2011).

5.1.3.4 Hb expression in surface glia cells is specific in carpet cells

Carpet cells have been shown to be a sub-population of the sub-perineural glia cells (Silies et al., 2007). However, I observed that a sub-perineural driver (NP2276 (Awasaki et al., 2008)) does drive reporter gene expression in brain sub-perineural glia, but not in carpet cells (data not shown). Additionally, using immunostaining with two different antibodies and two driver lines (VT038544 and VT038545) I could not detect *hb* expression in other surface glia cells, not in the eye disc nor in the larval brain. These data indicate that carpet cells are indeed a subtype different from other sub-perineural glia, since they express at least one specific marker (namely *hb*).

Microarrays have recently been used to reveal the transcriptome of adult blood-brain barrier surface glia (DeSalvo et al., 2014). Interestingly, the only overlap between the list of 50 highest expressed genes in these cells and my list of putative Hb target genes is the gene *fas2*. This supports the idea that carpet cells are a very specific type of cells, different from the rest of sub-perineural cells and that Hb can be defining this specificity. *hb* is not expressed in the other surface glia, and therefore it is also consistent that its targets are not expressed in the other surface glia. Additionally, it is also likely that the function of Hb in carpet cells is only performed during larval stages and probably not later during adult stages, when the transcriptome of surface glia has been analyzed.

In the analyzed brains, only one cell shows overlapping signal for Hb and for the pan-glial cell marker Repo (Figure 4.1.10). A staining overlap is more difficult to interpret in brain preparations, since these structures are more complex than the imaginal discs. Therefore, it cannot be excluded that this overlap could be an artifact.

5.1.3.5 Hb target genes can reveal its function in carpet cells

Finally, the study of putative Hb target genes has given new insights into the possible roles of this transcription factor in carpet cell development. Many of these genes have GO terms related to axon guidance and compound eye development, but also glia cell migration and development, Bolwig's organ morphogenesis and endoreduplication. Carpet cells are

polyploid cells, which is the result of endoreduplication process (Unhavaithaya and Orr-weaver, 2012). At least one of the putative Hb target genes has a function in endoreduplication (*archipelago* (Shcherbata et al., 2004)), which could be the cause of these enlarged cell nuclei. The Bolwig nerve is composed of the photoreceptor axons of the larval eye (also known as Bolwig organ), and these axons project through the optic stalk into the larval brain (Schmucker et al., 1997). It is known that axons can provide the necessary substrate for glia cells to migrate (Dearborn and Kunes, 2004) and it has been suggested that the Bolwig nerve could be involved in retinal glia migration and the development of the optic stalk (Schmucker et al., 1997). Also noteworthy is the fact that a large number of the identified Hb target genes are involved in the epidermal growth factor (EGF) pathway. This is a well-conserved pathway that has received a lot of interest due to its many roles in development and cancer (Gao et al., 2011; Sharma et al., 2007; Yewale et al., 2013). The activation of the EGF receptor (EGFR) by the binding of specific ligands initiates a signaling cascade (including MAPK phosphorylation pathway) that transmits information between cells during many different processes, including cell division, differentiation, cell survival and migration (reviewed in Shilo, 2003, 2005). Most of these roles of EGF pathway have also been documented as involved in *Drosophila* eye development (reviewed in Malartre, 2016). The list of Hb target genes up-regulated during eye-antennal imaginal disc development includes both positive regulators (*rhomboid*, *Star* and *CBP*) and negative regulators (*fasciclin2* and *sprouty*) of this pathway. I could show that at least *fasciclin2* and *rhomboid* are expressed in the region where carpet cells are located. However, this should be also checked at earlier stages, when carpet cell migration is more important, and including a glia cell or carpet cell specific molecular marker to confirm that these targets are expressed in *hb* positive cells during eye-antennal imaginal disc development. Multiple reports relating EGFR signaling with cell migration in different cancer types (e.g. Gao et al., 2011; Price et al., 1999) would also point in the direction of this process being possibly regulated by Hb in the carpet cells, in line with the results obtained in the loss of *hb* function and the *hb* misexpression experiments.

5.1.3.6 Hypothesis for Hb role in carpet cells and future work

At the moment, at least two different hypotheses are possible to explain the phenotypes I observe when the expression of *hb* is reduced or eliminated from carpet cells. On one hand, Hb could be necessary only to facilitate the migration of carpet cells through the optic stalk into the eye disc. It could be that carpet glia cells are present in their place of origin, but are not able to migrate into the eye disc because of the loss of *hb* expression. Many putative Hb

target genes in my dataset have GO terms related to migration and this could explain that wrapping glia cells that misexpress *hb* over migrate into the optic stalk (Figure 4.1.13). Since I do not observe over migration of perineural glia cells anterior to the morphogenetic furrow, which is a phenotype observed after carpet cell ablation (Silies et al., 2007), the cell body of the carpet cells that functions as scaffold for the basal perineural glia cells might still be present. It could be that the cell nucleus of the carpet cell remains at the base of the optic stalk and the cell margins are still able to grow into the eye disc. However, it is hard to imagine that the cells can grow and properly project their cell membranes to such far distance and still correctly coordinate the advance of the other perineural glia cells. As later, during pupal stages, carpet cells migrate back to the brain lamina (Edwards et al., 2012), the migration into the eye discs is probably an essential part of the proper function of carpet cells.

On the other hand, Hb could be necessary to specify carpet cell identity. In this case, in *hb* RNAi and *hb* knock-out discs the entire carpet cells would be missing. When *hb* is misexpressed in undifferentiated perineural glia cells, preliminary data shows that all the cell nuclei acquire carpet cell-like characteristics. While driving *hb* expression in differentiated wrapping glia cells (Mz97-Gal4 (Ito et al., 1995)) is probably too late to change their fate, thus this misexpression only in some way modifies their normal fate. Although it has been shown that the loss of carpet cells results in over migration of perineural glia cells, the corresponding experiments to ablate carpet cells were performed using *moody*>>*hid* lines (Silies et al., 2007) and hence the induction of cell death in *moody* expressing cells. This not only affects carpet cells but also other glia cells in the brain, thus the interpretation of such phenotypes may be problematic. Especially the effect that loss of *hb* function in glia cells has on blood-eye barrier integrity indicates that the carpet cells or at least a part of them are missing.

In order to distinguish between these two possibilities driver lines can be used to also mark the cell membrane of the carpet cells. This could be used in combination with the RNAi constructs to determine if the cell body of carpet cells is still present in the eye discs after *hb* expression is knocked-down in these cells. Alternatively, Moody antibodies (Bainton et al., 2005) can be used to visualize the cell body of the sub-perineural glia cells in the optic stalk and in the eye disc. In wild type animals, only carpet cells express moody in the optic stalk and eye discs (Silies et al., 2007; Figure 4.1.14A). The presence of Moody protein in the eye disc of animals that have *hb* expression knocked-down would indicate that at least their cell bodies are present in the eye discs.

Interestingly, while Hb role in anterior/posterior patterning seems to be conserved only in insects or arthropods (Pinnell et al., 2006; Schröder, 2003), its role in central nervous system development is conserved at least across all protostomes (Pinnell et al., 2006). One of the *hb* homologs known in mammals, *ikaros*, which also promotes early-born neuronal fate in mouse (Alsiö et al., 2013), has been shown to have a role in conferring identity to retinal progenitor cells (Elliott et al., 2008). Although this function in vertebrates is in neurons and not glia cells, this shows that the re-deployment of Hb in visual system development has happened more than once.

5.1.3.7 New Hb 3' regulatory region driving expression in nervous system

The location of the Hb regulatory regions that we have used to drive expression in the carpet cells (Pfeiffer et al., 2008) are accessible to DNA-binding proteins at embryo stages 9 and 10 and much less at stages 5, 11 and 14 (Li et al., 2008). Additionally, this DNA region does not seem to be bound by *bicoid* during early embryo stages (Supplementary Figure 1, (Li et al., 2008)). Early-born neuroblasts express *hb* specifically at embryo stages 9 and 10 (Grosskortenhaus et al., 2005). The overlap between the profiles of open chromatin and the selected region in the VT lines indicates that the regulatory region that drives expression in the carpet cells could be the same that drives expression in early-born neuroblasts. This regulatory region is located at the non-coding 3' end of the *hb* locus (Supplementary Figure 1). Although it is not frequent, other examples of genes with *cis*-regulatory elements in the 3' end are known, such as the gene *Pax6* (Griffin et al., 2002). This might be especially common for genes with the characteristics of *hb*, having multiple functions throughout development but with a small gene body (with only one coding exon and two introns, one of them less than 300 bp long) and located in a gene rich genomic region. A *cis*-regulatory region in the 5' end of the *hb* gene driving expression in neuroblasts and early-born neurons has also been already identified (Hirono et al., 2012). It is possible that these regions at the two ends of the *hb* locus interact with each other by looping mechanisms (Noordermeer and Duboule, 2013). Further analyses could be performed to clarify this. For instance, ATAC-seq could be applied (Buenrostro et al., 2013) to investigate if the 5' region, as well as the 3' region, is also accessible during eye-antennal imaginal disc development. Additionally, the 5' region could be cloned to a driver construct to study if it also drives expression in the carpet cells.

5.1.4 Conclusions and outlook

The detailed analysis of a putative new role of Hb in *Drosophila* visual system development not only confirmed that this transcription factor is expressed in the eye-antennal imaginal disc but it also revealed that it plays a crucial role in the development of a subtype of retinal glia cells. Different expression and functional analyses have helped to better understand the role that Hb plays in these cells. I have found that not only is Hb necessary for the proper development of carpet glia cells, but its presence is also necessary to ensure a proper separation between the hemolymph present in the body cavity and the retina.

The large cell body of these cells implies that any genes coding for cytosolic or membrane bound proteins present in these cells need to be highly expressed. However, the RNA levels of *hb* at LIII stage in eye-antennal imaginal discs is negligible, since it is only expressed in these two cells. At earlier stages these cells are not yet in the imaginal discs, and *hb* expression could have only been identified by studies focused on the optic stalk. Therefore, the identification of *hb* being expressed in carpet cells has only been possible through my method of target genes co-expression analysis. Moreover, I could show that the refined list of Hb target genes contains genes with GO terms highly specific for the putative function of Hb in carpet cells. Analogous analysis of some of the other identified transcription factors could also reveal new functions for these genes and find additional downstream target genes. A similar approach has already been successfully used to identify previously unknown key developmental regulators, for example a number of nuclear receptors involved in metamorphosis (Potier et al., 2014b) also in *Drosophila*.

All of this evidence demonstrates that the combination of high throughput transcript sequencing with ChIP-seq datasets enrichment analysis can reveal previously unknown factors and also their target genes, and therefore increase the number of connections of developmental GRNs. Other studies have searched for regulating transcription factors that were in the same co-expression clusters as its targets genes (Potier et al., 2014a). However, upstream orchestrators not necessarily have the same expression levels as their targets, and I could clearly show an instance of that with the example of Hb expression in carpet cells. Therefore the combination of ChIP-seq methods in RNA-seq co-expression analyses has proven to be a powerful tool to identify new developmental regulators that can complement other studies based on reverse genetics.

My data also contributes to the growing thought that most genetic networks are re-used in many different processes throughout development (Carroll, 2008). Reports of a genetic network governing muscle development in vertebrates being re-deployed during retinal differentiation in *Drosophila* have been described (Heanue et al., 1999). This could also help describe our surprising finding of Mef2 being maybe involved in eye-antennal imaginal disc development. My findings on the role of Hb in glia cell development also show the large pleiotropy of key developmental regulators, as it also has known functions in processes that would seem to have little in common with that, such as embryo segmentation (Lehmann and Nüsslein-Volhard, 1987; Nüsslein-Volhard and Wieschaus, 1980). This can be explained by the presence of multiple *cis*-regulatory elements in the locus of “toolkit” genes, such is the case of *Pax6*, which has at least 6 different *cis*-regulatory elements (Griffin et al., 2002), or the pair-rule gene even-skipped, which has *cis*-regulatory modules for at least 10 different transcription factors (Wilczynski and Furlong, 2010) or *hb* as I described in section 5.1.3.7. It is because many of the transcription factors regulating embryogenesis are also re-used at later developmental stages that the use of ChIP-seq experiments that had been performed on embryos has worked in my analysis. It could be suggested that ChIP-seq be performed for all “toolkit” genes on embryos, and subsequently, as I have presented here, only RNA-seq experiments would be necessary on the tissue and stages of interest to identify the specific targets expressed and thus the re-deployed genetic networks. Although target genes that are specific for the stage of interest and not during embryogenesis would not be identified with this method, the upstream transcription factors could be identified. Subsequent ChIP-seq experiments with the interesting transcription factors could then be performed to extend the list of target genes. Additional analyses on genome-wide chromatin accessibility (Buenrostro et al., 2013) or histone modifications (Pan et al., 2007; Pokholok et al., 2005) in the condition of interest could refine the list of target genes specific for that tissue and stage.

5.2 A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species

The reciprocal re-annotation of the genomes of the model species *D. melanogaster* and its closely related species *D. simulans* and *D. mauritiana* has been of foremost importance for the evolutionary studies I have performed during my Thesis (Torres-Oliva et al. in revision). I could show that this step is necessary for an unbiased, genome-wide analysis of differential gene expression across different species using RNA-seq technology for two main reasons: to prevent biases due to length differences between orthologous genes and to compare the largest number of orthologs when the genome assembly and genome annotation quality of the different used species differs.

By comparing my strategy to other available methods and by checking the results using quantitative RT-PCR on a number of genes, I could show that my strategy provides the most robust results. I could also show that RPKM-based methods, which could correct for the differences in length between orthologous genes, fail to do so when the expression levels are not homogeneous across the gene body (Table 4.2.3, Supplementary Figure 8). This introduces clear biases in the analysis of differential expression and leads to false positives.

Furthermore, the reciprocal re-annotation of the *D. simulans* and *D. mauritiana* genomes also allowed the interrogation of the maximum number of orthologous genes between these two species. In the published genome annotations, a large number of genes had been filtered out owing to incomplete genetic sequences. Remarkably, this included the lack of *ocelliless* (*oc*) annotation in the genome of *D. mauritiana* (Nolte et al., 2013) due to the fact that the 5' end of its locus contained many unspecific nucleotides. After the reciprocal re-annotation of *D. mauritiana* and *D. simulans* genomes, the longest orthologous region of the *oc* gene present in the two references was annotated and used as reference to map the RNA-seq reads. Thus I could detect that this gene is differentially expressed at late LIII stage eye-antennal imaginal discs between these species, which was later confirmed by *in situ* hybridization (see sections 4.4 and 5.4). Being able to examine the highest number of genes, of course, becomes especially relevant in studies that are performed to identify one or very few candidate genes.

In summary, the reciprocal re-annotation method I have developed is a necessary step to perform unbiased inter-species differential gene expression analyses prior to mapping RNA-seq reads to the species-specific references. Failing to do so, especially when working with non-model species, could reduce the number of genes to be analyzed and generate false positive results due to length differences between orthologous sequences caused by differences in the quality of the respective assembled genomes.

5.3 Gene expression divergence in closely related *Drosophila* species

5.3.1 Differential gene expression in closely related species

In a previous Discussion (section 5.1.1) I have described the high degree of gene expression conservation between the three closely related species of this study. In spite of that, a large number of significantly differentially expressed genes are also detected. Interestingly, while the number of differentially expressed genes between *D. melanogaster* and the other two sister species is largest at early stages (72h AEL) and lower later, the number of differentially expressed genes between *D. simulans* and *D. mauritiana* is low at 72h AEL and it rises at each subsequent stage (Figure 4.3.3). Although the sequencing type could have an effect on these results (see below, “Technical considerations”), differences in the developmental timing (heterochrony) (Olsson et al., 2010) could also be involved in this variation, as it has already been described to affect closely related species of *Drosophila* during other developmental processes, such as segmentation, without disturbing the developmental regulation of the underlying network (Kim et al., 2000). Although we checked the relative time point for each species when the morphogenetic furrow was located at the middle of the retinal field to dissect the mid LIII stage, we do not know the exact time point when retinal differentiation starts in *D. mauritiana* and *D. simulans*. It could be that in these two species it starts slightly earlier than in *D. melanogaster* and that could explain why at 72h AEL (which was the same dissection time for the three species) so many genes are differentially expressed. The enriched GO terms for the genes differentially up-regulated at 72h in *D. mauritiana* and *D. simulans* with respect to *D. melanogaster* are related to biological regulation and eye development, while the genes up-regulated in *D. melanogaster* at this stage are related to cell cycle process (data not shown, results from Elisa Buchberger Lab Rotation Protocol for her Master studies under my supervision). Additionally, the up-regulated genes in *D. mauritiana* and *D. simulans* are enriched for *cis*-regulatory elements bound by the transcription factor Nejire, which is a regulator of eye development (Kumar, 2004), while the genes up-regulated in *D. melanogaster* are enriched for *cis*-regulatory elements bound by Pannier, which during late LII represses eye development (Oros et al., 2010). Two hypotheses could explain these results. On the one hand, LIII could start earlier in *D. simulans* and *D. mauritiana* compared to *D. melanogaster* and therefore we have sequenced early LIII stage at 72h AEL in these species, when eye differentiation has already started (heterochrony). On the other hand, the timing could be

similar but, either the expression of the genes involved in the face determining network (regulated by Pannier) in *D. melanogaster* is up-regulated or the genes involved in eye differentiation are up-regulated in *D. simulans* and *D. mauritiana*. Both hypotheses are actually complementary, and they both would lead to the same resulting up-regulation of gene expression of one or the other set of genes. Interestingly, this observations could also be related to the fact that *D. melanogaster* has fewer ommatidia number than the other two species (Posnien et al., 2012). The fact that most genes in cluster 11, which are involved in generation of neurons and compound eye morphogenesis, are up-regulated in *D. mauritiana* compared to *D. melanogaster* (Figure 4.3.5C) would also favor the later hypothesis, and it could be caused by an extended LIII stage in *D. mauritiana* that led to higher overall expression of all retinal differentiation genes. Differences in the size of the eyes between *D. mauritiana* and *D. simulans* are only due to larger ommatidia facet (see project “**Eye size variation between closely related *Drosophila* species**”) (Arif et al., 2013; Posnien et al., 2012), and these differences are established only after 120h AEL (Arif et al., 2013), which corresponds with more genes being differentially expressed between these two species only in later stages (Figure 4.3.3).

Among these closely related species, the relationship between genetic network topology and divergence can also be studied. Most previous studies on this topic have been focused on individual metabolic pathways and have mostly analyzed divergence at the coding sequence level (e.g. Alvarez-Ponce et al., 2008; Davila-Velderrain et al., 2014). In general, these studies have not found a clear correlation between the position of a protein in a genetic network and the strength of positive selection acting on its sequence (Davila-Velderrain et al., 2014), although in some cases a negative correlation has been described, being that genes with more targets were much more conserved (Montanucci et al., 2011). Here, the combination of my developmental gene co-expression analysis with my data on inter-species expression divergence has allowed me to obtain information about the preferential location of genes with gene expression divergence in genetic networks (Figure 4.3.5). Although no clear conclusions can be drawn yet from the three analyzed networks, some of the genes that are differentially expressed between *D. melanogaster* and *D. mauritiana* at 96h AEL have rather central positions in these networks (e.g. *kay*, *Dl*, *Ret* or *syp*) and in some cases they are the genes that connect the different subnetwork (e.g. *Nsf2*, *Cdk4*, *asp* or *fn2*) (Figure 4.3.5A and B). In other recent studies, similar results have been described. In a study performed in yeast, the authors initially reported no clear correlation between network topology and gene expression divergence, although when they split the studied

network into smaller subnetworks, they could find that genes with more connections were slightly significantly more divergent at the gene expression level (Kopp and McIntyre, 2012). Another study in mammals found that gene expression divergence in the PI3K signaling cascade was mainly due to gene expression divergence of the two main regulators, mTOR and AKT2 (Monaco et al., 2015). My data could be further used to perform a genome-wide analysis on all the obtained co-expression clusters across the three *Drosophila* species in order to reveal if significant positive or negative correlation exists between gene connectivity and gene expression divergence. To my knowledge, such studies at the systems biology level have only been performed in yeast (Carlson et al., 2006) and they indicated that highly connected genes are less variable. However, my preliminary results (Figure 4.3.5) and the described studies in mammals might indicate that in higher organisms the situation is different (Monaco et al., 2015). Yet, to understand this problem at the mechanistic level, ASE studies can provide a better insight as they can distinguish the type of regulatory change that causes gene expression divergence between closely related species (Wittkopp et al., 2004).

5.3.2 Expression divergence in developing tissues could be mainly regulated in *trans*

In order to reveal whether gene expression divergence in developing tissues is more often caused by changes in the locus of the differentially expressed gene (due to changes in *cis*) or by changes in upstream factors that can be in any location in the genome and regulate more than one gene at the same time (changes in *trans*), I have analyzed the expression of species-specific alleles in F₁ hybrids of closely related species (Figure 2.3). My data clearly shows a larger percentage of genes being differentially expressed between species due to changes in *trans* (Figure 4.3.8). In contrast, previous ASE studies between *D. melanogaster* and *D. simulans* adult tissues have reported higher percentage of genes with divergent expression due to changes in their *cis*-regulatory elements (Graze et al., 2009; Landry et al., 2005; Wittkopp et al., 2004, 2008), while two studies including comparison to *D. sechellia* indicated slightly more genes with expression divergence explained by changes in *trans* (Coolon et al., 2014; McManus et al., 2010). However, the earlier of these studies were performed using pyrosequencing on only a small set of pre-selected genes, which thus can hardly be extrapolated to the complete transcriptome (Landry et al., 2005; Wittkopp et al., 2004, 2008). Also to note is the fact that the studies performed using RNA-seq technology have not included biological replicates (Coolon et al., 2014; McManus et al., 2010). It is

strongly discouraged to do RNA-seq analyses without at least 3 biological replicates (Hansen et al., 2011; Liu et al., 2014), since variance inherent in this method is large and differences in the number of mapped reads can be due to random sampling and not due to difference caused by the studied conditions. A sign of this bias could be in fact that in McManus et al. 2010 the authors identified as much as 78% of genes having significant differential expression between the two closely related species *D. melanogaster* and *D. sechellia*. In contrast, in my analysis, only between 11% and 19% of genes are significantly differentially expressed between species. Working without replicates could also influence the number of differential ASE detected, as a measure of the random variance for each gene within one species is not available, and any differences in the expression of one allele relative to the other allele could be detected as significant. This could result in a larger number of genes detected as having expression divergence because of changes in their *cis* regulation. In my analysis, I have used three biological replicates for each condition and I used the software that has been shown to perform best in multiple benchmark studies, namely DESeq2 (Love et al., 2014a) (Ching et al., 2014; Lin et al., 2016; Rapaport et al., 2013; Seyednasrollah et al., 2013), to execute the statistical analyses required to detect significantly differential expressed genes and alleles.

Additionally, an explanation for my data showing many more cases of gene expression divergence due to *trans* regulation is the fact that I study tissues that are undergoing developmental processes such as tissue patterning and organ differentiation. These processes are more commonly controlled by transcription factors (Carroll, 2001), and changes affecting one transcription factor can influence the expression levels of many target genes. An interesting result is that only the genes that change due to variation in their regulation in *trans* show, in some cases, different direction of change in the two different tissues I have studied (e.g. a gene can be higher in *D. melanogaster* in the eye-antennal imaginal disc but lower in *D. melanogaster* in the wing disc) (Figure 4.3.9). This actually supports the validity of the analysis pipeline, because a change on one *cis*-regulatory region of a target gene (which is bound specifically by one “toolkit” transcription factor) will affect the expression of that gene for the transcription factor that recognizes that region equally in all tissues (see Figure 5.1, green *cis*-regulatory element). However, *trans* regulation can vary according to the different transcription factors that are expressed in each tissue. Changes in the coding sequence of a transcription factor can make it bind with more or less affinity to the different *cis*-regulatory elements that it binds to, and its target genes can be different ones in each tissue.



- TF
- TF's eye disc CRE
- ◆ TF's wing disc CRE
- Target genes CRE
- * polymorphism

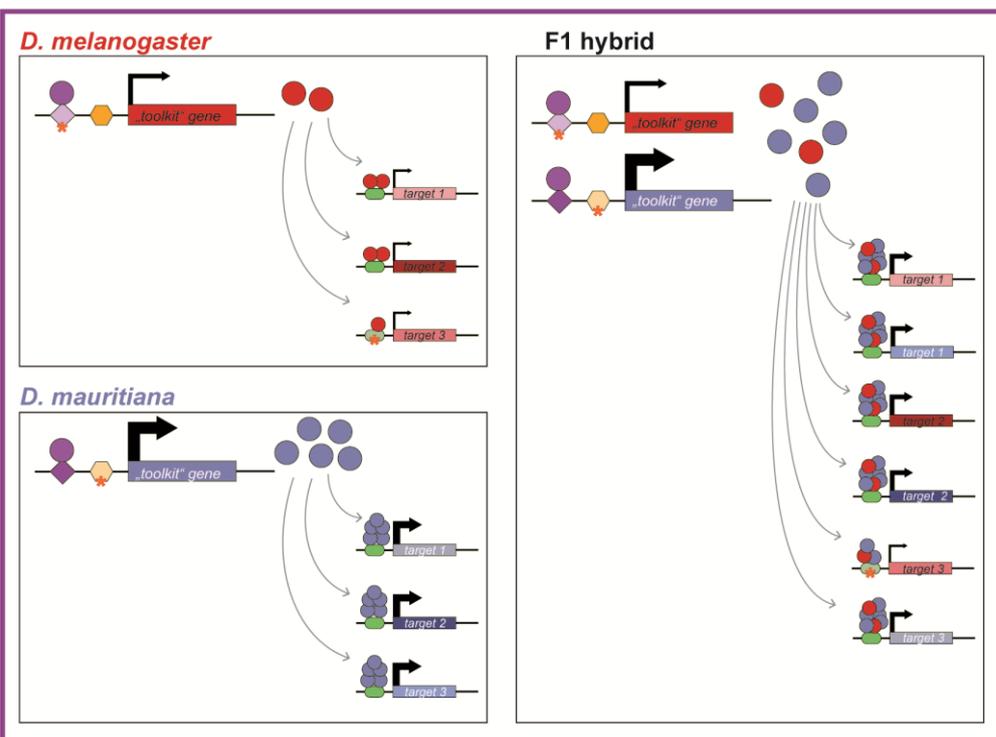
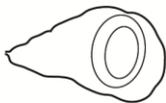
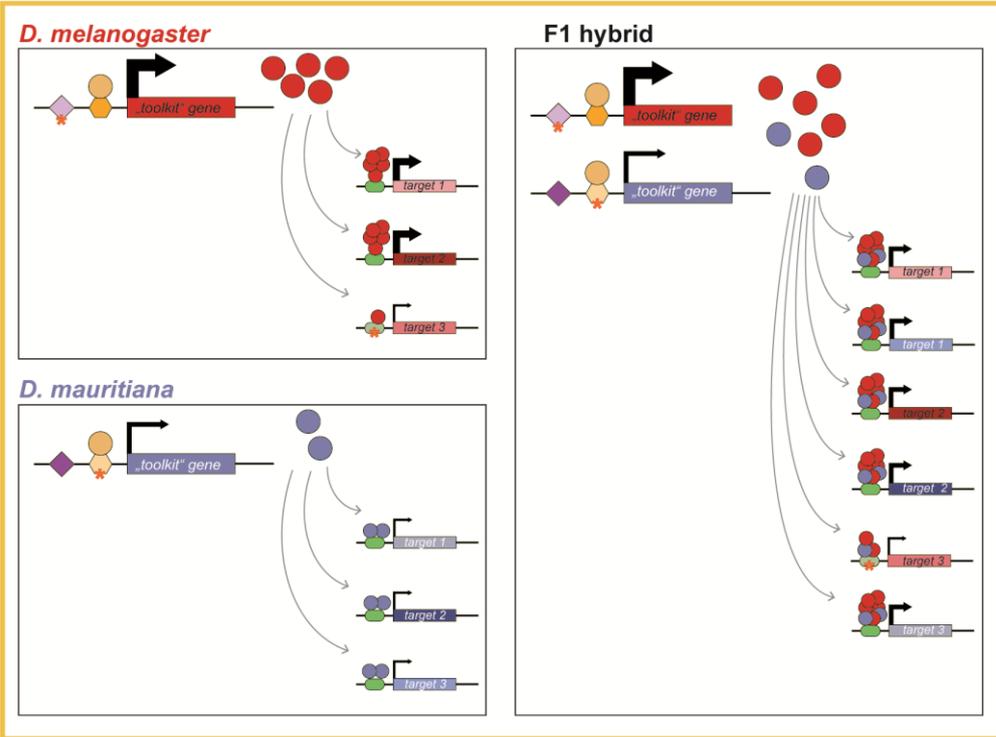


Figure 5.1. Changes in tissue-specific *cis*-regulatory elements of upstream transcription factors are more likely to produce gene expression divergence. “TF”: transcription factor; “CRE”: *cis*-regulatory element. Red horizontal bars represent *D. melanogaster* genes (left) or alleles (right). Blue horizontal bars represent *D. mauritiana* genes (left) or alleles (right). Thickness of the black arrows represents the different expression levels. The modular nature of CREs makes it possible for one gene (in this case a “toolkit” transcription factor) to be differentially regulated in different tissues. In this scheme, the lila CRE represents a region active in the wing disc, and the yellow CRE represents a region active in eye-antennal imaginal disc. The yellow CRE has a mutation in *D. mauritiana* with respect to *D. melanogaster* that results in a lower expression of the transcription factor in the eye-antennal imaginal disc. Therefore, all the target genes have lower expression in the eye-antennal imaginal disc of *D. mauritiana*. In the hybrid, though, all target alleles have equal expression because the two CREs regulating the two alleles of the transcription factor are present in the cells. The other way around, the lila CRE has a mutation in *D. melanogaster* with respect to *D. mauritiana* that results in a lower expression of the transcription factor in the wing disc. Thus, all the target genes have lower expression in this tissue in *D. mauritiana*, but the alleles have equal expression in the hybrid. Alternatively, a mutation in the CRE controlling the expression of a target gene (green CRE) affects the expression of that gene in the parent and in the hybrid, and equally in all tissues.

5.3.2.1 Evolutionary implications of the high number of divergent genes due to changes in *trans*

Due to the larger pleiotropic effects that *trans* regulatory changes can have, as they can affect a large number of target genes, it is generally thought that *cis* changes are favored as main cause of gene expression divergence between species (Carroll, 2008; Warnefors and Kaessmann, 2013). However, my results clearly show a higher percentage of genes showing divergent expression due to changes in the upstream factors that regulate their expression. This variation in *trans* can be explained by changes that affect the coding regions of the regulating transcription factors or by changes in the regulation of these transcription factors, i.e. in the *cis*-regulatory elements of “toolkit” genes. My hypothesis is that these differences are mostly due to differences in the regulation of the expression of the transcription factor genes and not so much due to coding differences. Especially due to the fact that most components of the genetic “toolkit” are highly conserved at the protein sequence level (Halder et al., 1995; McGinnis et al., 1990). The scheme in Figure 5.1 describes this possibility. Transcription factors from the genetic “toolkit” have been shown to present usually several different *cis*-regulatory elements (Davidson, 2001; Griffin et al., 2002; Stanojevic et al., 1991). As I previously described, these transcription factors and the genetic networks that they regulate are often co-opted in different cell types and at different stages to perform different functions. These different *cis*-regulatory elements can allow this, as they are highly modular and one can evolve without disturbing the others (Figure 5.1). To test this hypothesis, the putative upstream regulators of genes found to be divergent due to changes in *trans* could be searched (for instance using i-cisTarget

(Herrmann et al., 2012)) and the expression levels of each upstream regulators could be checked in the two parental species to investigate if they are differentially expressed due to changes in *cis* regulation. For instance, some of the genes that have expression divergence due to changes in *cis* are enriched for GO terms related to biological regulation, and could include some of the “toolkit” transcription factors that change in expression and affect other target genes in *trans*.

Examples have already been described of *cis*-regulatory evolution of genetic “toolkit” genes giving rise to morphological divergence (Belting et al., 1998; Shapiro et al., 2004; Sucena and Stern, 2000). In most cases this is indeed by the presence of multiple *cis*-regulatory elements regulating the upstream transcription factor. In this case, a transcription factor that works correctly in one tissue, by the addition or modification of a regulatory region that could create a new function in a different tissue, does not disturb the function of the transcription factor in the original tissue. Sequence turnover of *cis*-regulatory elements is also much higher than that of the coding sequences, as they have less selective constraints (Prud'homme et al., 2007) and expression differences can have fewer pleiotropic deleterious effects. Therefore, it is more likely that between closely related species, more changes in the non-coding region of genes are present. My data allows to further test whether at early steps of speciation, among very closely related species, expression divergence of transcription factors governing the expression of many other target genes is the main cause of expression divergence of genes involved in developmental and patterning processes.

However, changes in the coding regions of the transcription factors cannot be completely ruled out. It has been shown that, although the “toolkit” genes are extremely well conserved because genes from distantly related taxa can be exchanged and still perform most of the original functions (Halder et al., 1995), in some cases great divergence at the amino acid level can be observed. The protein domain that binds DNA is usually highly conserved, but the rest of the protein is not subject to such strong negative selection (Wagner and Lynch, 2008). That is why in some experiments where authors have replaced the ortholog of one species by another, they could reproduce some phenotypes but not all (Ranganayakulu et al., 1998), as the proteins conserve the specific DNA binding domain and specificity for some *cis*-regulatory elements but not necessarily for all the original ones. This divergence in protein structure while conserving functionality can be favored by the presence of multiple *cis*-regulatory elements in the regulatory region of these regulatory genes. While conserving the original regulatory region, a “toolkit” gene can be re-used in

another tissue to perform a new function. Small changes that do not entirely disturb the original function, such as changes that code for modification in the protein's unstructured regions or by means of alternative splicing, can take place to accommodate the new function.

Finally, it is also worth to comment on the fact that between *D. simulans* and *D. mauritiana*, a large number of genes show compensatory regulation (Figure 4.3.8). This means that the species-specific alleles in the hybrid show significant differential expression, but in the parents, the orthologs are not differentially expressed. This is thought to be caused by co-evolution of regulatory elements within each parental species that do not give rise to actual expression differences. However, in the hybrid, epistatic effects of one allele over the other become evident and these differences are expressed (Goncalves et al., 2012; Landry et al., 2005; Tirosh et al., 2009). In closely related species such as the ones of this study, it may reflect changes that are not yet fixed in the population. These changes, as the expression divergence caused by changes in the *cis* region of genes, do not change direction in the different tissues (Figure 4.3.9).

5.3.3 Technical consideration

5.3.3.1 Inter-species RNA-seq using different sequencing types

The reciprocal re-annotation method has allowed the comparison of gene expression levels of 13,457 genes across the three closely related *Drosophila* species (Torres-Oliva et al. in revision). It is to note, however, that some bias could have been introduced by the use of different sequencing types. Although the multidimensional scaling plot of the normalized datasets indicates that the sequencing type (single-end 50 bp or paired-end 100 bp) is not the main factor separating the data (Figure 4.3.1), still the two samples from *D. mauritiana* 120h AEL that were sequenced with the two methods are slightly separated by dimension 2. This possible bias could explain that I had to re-cluster some of the clusters because all genes that were up-regulated at 96h and 120h AEL were grouped together in only two clusters (Figure 4.3.2). In order to better compare expression levels from the two sequencing types, I split the 100 bp reads by half and used the right and left paired reads together as single end reads. This improved the comparability of read counts between samples. But the samples coming from paired-end 100 bp sequencing had therefore more than 4 times more reads than the single-end 50 bp, as the throughput from this type of sequencing is much larger. It is clear that normalization plays a pivotal role in this kind of

analysis (Dillies et al., 2012). The latest version of the DESeq2 software (Love et al., 2014a), for instance, reports that they have implemented a tool to include different sequencing types in the RNA-seq comparison analysis as a condition to mask. However, the different sequencing types should be split into different conditions. In my study, I used paired-end sequencing for all 72h AEL *D. simulans* and *D. mauritiana* samples and single-end sequencing for the 72h AEL *D. melanogaster* samples. Therefore, although the normalization applied by DESeq2 helps to equalize the dataset, some indications show that it cannot fully normalize all read counts. For instance, many more genes appear as significantly higher expressed in *D. simulans* and *D. mauritiana* with respect to *D. melanogaster* at this stage (Figure 4.3.3). A closer look at the genes that had such high expression revealed that they are cuticle, myosin, ribosomal proteins and also many are unknown proteins. These genes could have very high expression levels that cannot be properly normalized in relation to the expression levels of the rest of the genes and therefore appear as up-regulated in the species with more total number of reads due to originating from paired-end 100 bp sequencing. Some of these genes that appear expressed at 72h AEL could be due to contamination during the dissection procedure, as eye-antennal imaginal discs at this stage are extremely small and their dissection is much more difficult than at later stages. Despite this, and after acknowledging and overcoming this problem by re-clustering the genes which are up-regulated at later stages, the analysis has produced unbiased results, as shown by the similarity of the biological processes of the different clusters to the results obtained with only *D. melanogaster*. As I recognized this problem, I proceeded to use the same sequencing type in all subsequent analyses, and all samples used for ASE analysis were sequenced using single-end 50 bp reads to prevent this type of bias. Probably a good addition to this analysis would be the dissection and sequencing of one biological replicate of *D. mauritiana* and *D. simulans* 72h AEL eye-antennal imaginal disc using single-end 50 bp. These samples could be used to properly normalize the other replicates that were sequenced using paired-end 100 bp reads. However, as more and more RNA-seq experiments are performed by different groups that could be interesting for other researchers to compare, it will be important to develop normalization methods that can account for these differences.

5.3.3.2 ASE analysis between very closely related species

I was also able to show that ASE analyses can be performed even between species so closely related as *D. simulans* and *D. mauritiana* (0.5 mya divergence time). However, an

elaborate preparation of the references is required prior to mapping the hybrid reads. The coding sequences of genes are more conserved across species, and therefore are more reliable to identify orthologous sequences. Yet sequence conservation is precisely not required for ASE analyses, as only polymorphic positions can be used to recognize the origin of allele-specific reads. This is an intricate trade-off, as the more distantly related the species are, the more polymorphisms they have between orthologs; however, the less likely and complicated it becomes to obtain viable hybrids when crossing them. In my case, I could cross *D. simulans* and *D. mauritiana* without problems but a large number of genes had zero or very few polymorphisms in the coding region of orthologous genes. *D. melanogaster* could be crossed with *D. mauritiana* but not with the strain we used of *D. simulans* (YVF).

In cases where references are missing for any of the species or where the available reference for one of the species is of superior quality than the others, some authors have proposed to perform *in silico* assembly of the genomic references of these later species using the species of better quality as scaffold (Munger et al., 2014; Shen et al., 2013), in my case *D. melanogaster*. Yet this is not an ideal solution, since it can be especially problematic to resolve insertions and deletions (INDELs). For the three species I used in my study (*D. melanogaster*, *D. simulans* and *D. mauritiana*) very recent publications (dos Santos et al., 2014; Hu et al., 2013; Nolte et al., 2013 respectively) have provided high quality genomes, and therefore I decided to use them. However, because these sequenced strains are not the ones I used in my analysis, I considered it necessary to use strain-specific genomic reads to perform *in silico* replacement of polymorphisms between the used strain and the published strain for each species (Satya et al., 2012; Shen et al., 2013). After this correction, the number of polymorphisms between *D. melanogaster* and *D. mauritiana* was large enough to correctly identify ASE. However, this strategy was not sufficient to unequivocally identify the species of origin of the reads in the hybrids of *D. simulans* and *D. mauritiana* and an additional *in silico* transcriptome assembly using the parental RNA-seq reads was necessary (Satya et al., 2012). The annotation of full transcripts (including UTRs) instead of only coding sequence also greatly increased the number of polymorphisms between orthologs and thus the specificity of the ASE analysis.

To verify the efficiency of these strategies, I analyzed the total number of polymorphisms between orthologous pairs before and after the process of *in silico* polymorphism replacement. Additionally, I investigated the allele specific expression of all mitochondrial genes, which should only show expression of the allele coming from the species that contributed as female in the cross (McManus et al., 2010; Reilly and Thomas, 1980). The

error rate for species-specific allele assignment can be measured as the proportion of reads from mitochondrial genes that are misassigned to the *D. mauritiana* reference, which was the male in both crosses. With my analysis pipeline I obtained less than 0.3% error rate in the analysis of *D. melanogaster* and *D. mauritiana*, which is less than was reported in a previous ASE analysis between *D. melanogaster* and *D. sechelia* (McManus et al., 2010). Between the very closely related species the rate was higher, but still only 2.5%. I could show that, in mitochondrial genes, this is mainly due to a single polymorphism that had not been correctly replaced (Supplementary Figure 11) and this probably does not change the overall results.

5.3.4 Conclusions and outlook

The study of the biological processes and regulators of the differentially expressed genes between these closely related species has indicated that the underlying processes could be slightly different. For instance, *D. melanogaster* seems to start its retinal differentiation processes later than *D. mauritiana* and *D. simulans*. This could easily lead to differences in the overall final size of the respective adult organs (Amore and Casares, 2010; Kim et al., 2000). Analysis of the precise timing of the start of retinal differentiation in *D. simulans* and *D. mauritiana* will be necessary in order to determine whether heterochrony plays a role in setting the differences in relative head sizes between these two species and *D. melanogaster*.

The genes with expression divergence due to *trans* regulation and which are higher expressed in *D. melanogaster* with respect to *D. mauritiana* show a very high enrichment (NES=9.48) for Pannier binding sites. Additionally, a cluster of genes with higher expression in *D. melanogaster* compared to *D. simulans* and *D. mauritiana* (Figure 4.3.4, green cluster) shows also a great enrichment (NES=7.17) for the GATA motif in their *cis*-regulatory regions, to which Pannier is known to bind to (Romain et al., 1993). Therefore, Pannier is a strong candidate for being one of the transcription factors that has evolved and caused many of its target genes to modify their expression dynamics, i.e. expression divergence due to changes in *trans*. As previously mentioned, Pannier is known to have a role in defining face cuticle and repressing eye determination genes, as it has been shown that removal of *pannier* expression leads to ectopic enlargement of the eye region (Singh and Choi, 2003) and its overexpression suppresses eye fate (Oros et al., 2010). The student MSc. Elisa Buchberger has started her doctoral project to further investigate the putative evolution of Pannier to regulate the differences in head cuticle that can be observed between *D. melanogaster* and *D. mauritiana*. The coding sequence of the two orthologs is

highly conserved, although a few amino acid changes are present that could influence the binding affinity to the regulatory regions of its targets or the affinity to bind to other co-factors. A second likely hypothesis is that changes of the gene regulation are causing the differences in the expression of its target genes. Unfortunately, the expression levels of this gene in my RNA-seq data are very variable within biological replicates and it cannot be confidently asserted if *pannier* is differentially expressed between *D. melanogaster* and *D. mauritiana* during eye-antennal imaginal disc development. Interestingly, four different isoforms are reported to be expressed from the *pannier* locus, and already two of them have been found to be differentially expressed in the wing disc (Fromental-Ramain et al., 2008) and during embryogenesis (Minakhina et al., 2011). Again, the RNA-seq data that we currently have is not sufficient to correctly reveal the expression levels of the different isoforms in the eye-antennal imaginal disc, probably also due to the different expression domains of this gene. Apart from its function to promote head cuticle specific fate by repressing eye specific fate (Oros et al., 2010), earlier it is also involved in determining the dorsal-ventral axis of the eye disc by activating a cascade that includes Wg and the Iroquois Complex (Singh and Choi, 2003; Singh et al., 2005). Quantitative PCR analyses could be performed to determine if the expression levels of *pannier* vary between *D. mauritiana* and *D. melanogaster* during eye-antennal imaginal disc development. Moreover, available *D. melanogaster* lines that lack a full *pannier* locus (deficiency lines) will be crossed to wild type *D. mauritiana* flies to generate heterozygotes that have only a *D. mauritiana* functional copy of *pannier*. The measurement of the face size of these individuals compared to wild types flies can indicate if *pannier* is indeed the gene that has evolved to generate differences in the size of the face of these two species.

To my knowledge, this is the first time that a transcription factor with divergent expression levels has been identified by the expression divergence of its target genes. This has already been described as a challenging task mainly in the case of *cis*-regulatory evolution, as this variation is more difficult to identify than coding sequence divergence (Coolon and Wittkopp, 2013). Again, the used strategy of combining ChIP-seq datasets with our RNA-seq data using i-cisTarget (Herrmann et al., 2012) has provided a link between differential expression of a large number of target genes and their putative upstream regulator(s), despite the fact that the ChIP-seq experiment had been performed in a different stage (Junion et al., 2012). My results indicate that most of the changes in gene expression present between closely related *Drosophila* species are caused by differences in the expression of their upstream regulators. Current research seems to indicate that, especially

during developmental processes, these regulators are mostly members of the genetic “toolkit” of developmental genes. Although this could be due to the fact that these genes are also more extensively studied, it would be of great interest to perform additional comparable ChIP-seq experiments using all the known transcription factors that fall into this category (Rokas, 2008). There is obviously not a single way how evolution can work to define new morphologies (Alonso and Wilkins, 2005; Wagner and Lynch, 2008), rather any change that is beneficial for an organism’s adaptation can be fixed. There have been reports of all kind of changes: either a coding change in a terminal gene related to physiology (Hilscher et al., 2009), *cis*-regulatory changes of this gene (Galant et al., 2002; Manceau et al., 2011), a change of the conformation of the transcription factor that regulates several downstream genes (Löhr and Pick, 2005) or the complete re-deployment of a genetic network by addition of a new *cis*-regulatory element on an upstream regulator (Belting et al., 1998). However, my work brings a new unexpected result to the standing controversy of whether changes in *cis* or *trans* are favored by natural selection to shape the different phenotypes (Carroll, 2008; Prud’homme et al., 2007; Wagner and Lynch, 2008; Wittkopp et al., 2004). That is, that changes in *trans* are not so uncommon as it was thought, and rather are the predominant cause of gene expression divergence during developmental processes that pattern different tissues.

5.4 Eye size variation in two closely related *Drosophila* species

We have applied a combination of unbiased methods such as QTL mapping and differential expression analysis to identify the genetic basis of the differences in ommatidia size between the two closely related species *D. simulans* YVF and *D. mauritiana* TAM16 (Arif et al., 2013; Hilbrant et al., 2014; Posnien et al., 2012). We could reduce the number of candidates to only 14 genes. Additionally, the application of different molecular analyses further reduced this list to only 5 putative genes, namely *Es2*, *Glutamate-cysteine ligase catalytic subunit (Gcl)*, *ocelliless (oc)*, *Serine Protease Immune Response Integrator (spirit)* and *Tyramine β hydroxylase (Tbb)*. The analysis of coding sequence divergence in all genes present in the QTL region did not show clear evidence for any additional candidate showing important structural differences. Finally, the use of prior functional knowledge pinpointed the gene *oc* as the most likely candidate gene to be responsible for the observed differences in ommatidia size.

5.4.1 Identification of candidate genes to regulate eye size differences between closely related species

QTL analysis has been already successfully used in micro-evo-devo studies to identify evolved genes that cause morphological variation (e.g. Chan et al., 2010; Hilscher et al., 2009; Steiner et al., 2007). However, even when a large number of visible and molecular markers are available to distinguish the genomic regions of the different studied species, this method most often identifies regions containing several genes. The identification of the evolved gene is generally the limiting step of this type of studies. Here we have demonstrated that the use of next generation sequencing of expressed transcripts can be used to reduce the list of candidate genes. With the analysis of differential gene expression based on RNA-seq data we could rule out 83% of the genes present in the QTL region, resulting in a list of 14 candidate genes that could be individually tested for their involvement in eye development using different molecular techniques established in *D. melanogaster*.

We analyzed the expression domains of these 14 genes in the eye-antennal imaginal discs of *D. simulans* and *D. mauritiana* and also the effect of their knock-down in eye-specific regions in the model species *D. melanogaster* using RNAi (Table 4.4.1). Only 5 of these genes showed that they could play a role in eye development. *Es2* is ubiquitously expressed in the

eye-antennal imaginal disc in *D. simulans*, while it seems to lack expression in the posterior end of the eye region in *D. mauritiana*, and its knock-down gave a phenotype both at 25°C and 28°C. *spirit* is also ubiquitously expressed, spatially equally in the two species and the repression of its transcripts in the eye region also gave rise to morphological phenotype at 25° and 28°C. *Tbb* and *Gclc* have both a smaller expression domain posterior to the morphogenetic furrow and resulted in weak phenotypes when their expression was knocked-down in the eye region, only at 28°C. *oc* is expressed in the ocellar region of the eye-antennal imaginal disc, but also in the posterior region of the eye field. This expression domain was clearly wider in *D. mauritiana* TAM16 compared to *D. simulans* YVF (Table 4.4.1, Supplementary Figure 12), and further analyses have shown that this expression also starts earlier in development in the former species (Dr. Isabel Almudi, unpublished). This gene also resulted in the strongest phenotypes when its expression was knocked-down in the eye region. Due to the fact that *oc* is also the only one of the candidates that has described roles in eye development in *Drosophila* (Royet and Finkelstein, 1995; Tahayato et al., 2003) this is currently our main candidate to have evolved to generate the observed ommatidia size differences between *D. mauritiana* TAM16 and *D. simulans* YVF.

Thus, combining evolutionary studies with developmental knowledge allows the identification of genes responsible for morphological variation. Evo-devo can therefore be used to reduce the number of candidate genes, but the fact that a gene is not well studied does not mean that it may not contribute to the evolution of phenotypes. For instance, recent studies have used a similar approach to identify genes responsible for evolution of genitalia morphology in closely related *Drosophila* species (Tanaka et al., 2015). From the 6 genes they have identified as putative candidates, none of them had been previously described to play a role in the development of this trait. Therefore, the use of unbiased methods like RNA-seq to identify all genes that are differentially expressed, regardless of prior knowledge, is a useful step to identify candidate genes. If *oc* does not prove to be the evolved gene in following experiments, we will readily test the high ranked candidate *Es2* and *spirit*, for which no role in eye development has been previously described in *D. melanogaster*.

5.4.2 *ocelliless* is the main candidate underlying ommatidia size variation

The *oc* gene (also known as *orthodenticle* (*otd*)) is a well-studied homeobox transcription factor that is involved in early embryogenesis, where it has a role in head and brain

segmentation (Finkelstein and Boncinelli, 1994; Tallafuss and Bally-Cuif, 2002), and also in eye and brain development (Finkelstein et al., 1990). It has been described that *oc* plays a role in the early specification of the optic lobes and, in the adult, in the development of the inter-neurons of the visual system (Schmidt-Ott et al., 1994, 1995). *oc* is also expressed in the photoreceptors of all the visual systems of the fly: the larval eye and the adult compound eye and ocelli (Finkelstein et al., 1990; Vandendries et al., 1996). In the compound eyes, *oc* is necessary for the proper formation of the rhabdomeres and also for their proper subtype specification (Fichelson et al., 2012; Mishra et al., 2010; Tahayato et al., 2003; Vandendries et al., 1996).

Although a role in determining ommatidia size has not been described for *oc*, one of its known mutant alleles, *otd^{mi}*, affects the morphology of the photoreceptors, specifically by causing abnormalities in the shape of the rhabdomeres and even causing rhabdomere duplication (Vandendries et al., 1996). These observations make this gene, and especially the regulatory region affected by this allele, a good candidate to regulate the differences in ommatidia size between *D. mauritiana* and *D. simulans*.

Interestingly, *oc* function in *rhodopsin* expression can also give a hint of its possible evolution in these two species. Here I will describe the known link between *oc* expression and specific subtypes of Rhodopsins and photoreceptors. Additionally, I will report what is known about these ommatidia subtypes in the studied species. Although we have no data relating *rhodopsin* expression with ommatidia size, I propose how further investigations in this direction could be pursued.

Each ommatidium contains 8 light-sensitive photoreceptors, called R1-R8, which form regular clusters with 6 outer photoreceptors (R1-R6) and two inner photoreceptors, with R7 being located on top of R8 (Wolff and Ready, 1993). Each photoreceptor forms extensive membrane foldings, called rhabdomeres, which contain the Rhodopsins that gather the incoming light. There are 6 different Rhodopsin types, Rh1-Rh6 (Hardy, 1985; Zuker et al., 1985). All outer photoreceptors express *rh1* and thus it is the *rhodopsins* expressed in the inner photoreceptors that define the ommatidia type. In “yellow” ommatidia, R7 expresses *rh4* (which detects ultra-violet light) and R8 expresses *rh6* (which can detect green light); in “pale” ommatidia, R7 expresses *rh3* (which also detects ultra-violet light) and R8 expresses *rh5* (which detect blue light) (Bell et al., 2007; Chou et al., 1999; Papatsenko et al., 1997; Yamaguchi et al., 2010); additional ommatidia types have been described, like the “dorsal rim area” (DRA) ommatidia, where both R7 and R8 express *rh3* (Fortini and Rubin, 1990), the “dorsal yellow” type, where R7 expresses both

rh3 and *rh4* (Mazzoni et al., 2008) and R8 expressed *rh6*, and finally the “odd-coupled” ommatidia type, where R7 expresses *rh3* and R8 expresses *rh6* (Wernet et al., 2006).

It has been shown that *oc* specifies *rh3* and *rh5* expression and represses the expression of *rh6* (Tahayato et al., 2003). The promoter region of these three genes contains the binding site (TAATCC) for the homeodomain of *oc*. The same protein region acts to activate *rh3* and *rh5* and to repress *rh6* in “pale” ommatidia (McDonald et al., 2010). Posnien et al. 2012 showed that in *D. mauritiana* TAM16 eyes, the proportion of ommatidia expressing *rh3* is much higher than in *D. simulans* YVF. Conversely, the proportion of *rh6* expression in *D. simulans* is higher than in *D. mauritiana* TAM16. Unfortunately, in this study the proportion of *rh5* could not be analyzed. Hilbrant et al. 2014 performed a similar analysis to study the proportion of *rhodopsin* expression in different *Drosophila* species, including another *D. simulans* strain (Zom4). Although they did not include *D. simulans* YVF in the analysis, the proportion of *rh5* expression compared to *rh6* expression in *D. mauritiana* TAM16 was significantly higher than in all the other studied species. These results would favor the notion that *D. mauritiana* TAM16 has a significantly higher expression of both *rh3* and *rh5* than *D. simulans* YVF, while the later has significantly more ommatidia expressing *rh6*. This corresponds with the fact that *oc*, for which we have identified a significantly higher expression in *D. mauritiana* TAM16 compared to *D. simulans* YVF, precisely activates *rh3* and *rh5* and represses *rh6*.

In our RNA-seq dataset we have detected higher expression of *oc* in *D. mauritiana* compared to *D. simulans* at late LIII stage eye-antennal imaginal discs. Differences in eye size between *D. mauritiana* TAM16 and *D. simulans* YVF, unlike differences in face size, cannot yet be observed at this stage (Arif et al., 2013). Likewise, photoreceptors do not express any *rhodopsins* yet, since *rhodopsin* expression does not start until mid-pupal stage (Earl and Britt, 2006). Although we detect *oc* expression during LIII stage, it is only the expression of the *spineless* (*ss*) gene in R7 photoreceptors during mid-pupation that initiates the expression of the different *rhodopsins* (Wernet et al., 2006). Although it is thought that the determination of the different ommatidia subtypes is stochastic (about 70% are “yellow” and 30% are “pale” (Wernet et al., 2006)), it has been shown that these proportions vary among different species, and a certain level of plasticity exists (Posnien et al., 2012). The higher presence of *oc* in *D. mauritiana* TAM16 could account for a higher percentage of ommatidia expressing *rh3* and *rh5*, as *oc* expression throughout the retinal field can be already established during LIII stage, and influence later the percentage of different ommatidia type according to *rhodopsin* expression under *ss* expression in pupal stage.

Additionally, in *D. mauritiana* TAM16 eyes are especially enlarged in the dorsal region, and that has been thought to be caused by the presence of the DRA ommatidia type, which express *rh3* in the two photoreceptors (Posnien et al., 2012). Our *in situ* hybridization analysis showed that this is the area where more *oc* expression is detected in *D. mauritiana* TAM16 compared to *D. simulans* YVF, which could also explain the higher proportion of DRA ommatidia in this region.

The transcription factor Pph13 has been identified to counteract Oc activity (Mishra et al., 2010). Both genes are necessary for the correct biogenesis of rhabdomeres and Pph13 is necessary for *rh2* and *rh6* expression. I asked whether this gene could have also evolved and could also contribute to establish the inter-species differences in *rhodopsin* expression. However, *pph13* gene is located on the 2L chromosome and therefore does not correspond to any of the identified QTL regions and it is not significantly differentially expressed between these two species at 120h AEL (data not shown).

5.4.3 Ommatidia structure in *D. simulans* and *D. mauritiana*

In the previous section I have discussed that the presence of more “pale” and/or “DRA” ommatidia could be due to a higher *oc* expression in *D. mauritiana* TAM16. However, this idea does not explain why the ommatidia in this species are larger than in *D. simulans* YVF. It is possible that the presence of one Rhodopsin type is related to the size and the shape of the ommatidia, or that Oc affects other target genes that have a role in defining ommatidia size and/or shape. Posnien et al. 2012 measured ommatidia facet only in the 5 central ommatidia, thus it is not clear if this size difference accounts for the total compound eye size difference. If indeed photoreceptors expressing *rh3* and/or *rh5* are larger than those expressing *rh2*, *rh4* and/or *rh6*, it could be that in *D. mauritiana* TAM16 significantly more *rh3*-/*rh5*-expressing photoreceptors are present in the ommatidia located in the center of the eye. We have confirmed that the tested protocol for *Drosophila* head optical sectioning (Smolla et al., 2014) can be combined with fluorescent markers and the signal is not lost during the process of bleaching and clearing (unpublished results). Thus cell marking experiments, using for example Phalloidin staining together with α -Rh5 antibodies to detect photoreceptor type, could be employed to measure photoreceptor cell size. A comprehensive comparison of the different photoreceptor types’ size could indicate if differential *rhodopsin* expression correlates with certain differences in ommatidia size.

5.4.3.1 Evolutionary and functional implications of different ommatidia morphologies

Optical sectioning of *D. mauritiana* and *D. simulans* adult heads allowed a precise measurement of various ommatidia features. By measuring lens and pseudocone width of the central ommatidia, I could confirm the results from Posnien et al. 2012 and Arif et al. 2013, showing that ommatidia of *D. mauritiana* TAM16 are larger than those of *D. simulans* YVF on their more distal part. Ommatidia length was not significantly different between the species, although the mean value in *D. mauritiana* was larger than for *D. simulans*. The number of ommatidia in the analyzed eye midline was larger in *D. simulans* YVF. These two species have similar total number of ommatidia (Posnien et al., 2012), but *D. mauritiana* TAM16 has an enlarged dorsal region, with also more ommatidia in this eye region. Thus it is necessary that *D. simulans* YVF has more ommatidia in other regions to compensate this difference and still have similar total number of ommatidia.

The optical sections obtained by clearing and imaging with confocal laser scanning microscope (Smolla et al., 2014) can also be used to generate 3D reconstructions of the complete compound eyes (Figure 4.4.3A). This can be used to measure ommatidia volume and analyze ommatidia shape. For instance, it has been shown that more conical ommatidia, with larger aperture at the distal part and narrower at the base, increases the amount of light received by the rhabdomeres, and are thus found more commonly in species adapted to darker environments (Land et al., 1999). Measurements of inter-ommatidial angle could also be compared between the studied species, since this feature can have a direct impact on visual acuity (Hecht and Wolf, 1929; Warrant and McIntyre, 1993). Larger facet diameter can increase light sensitivity, but usually at the expense of decreasing image resolution. However, if the distance between the lens and the retina (pseudocone height) is enlarged, for example by a decrease in the inter-ommatidial angle, image resolution can be increased (Horridge, 1978; Warrant and McIntyre, 1993). It has been shown, for instance, that animals that have to fly through dense foliage concentrate ommatidia with small inter-ommatidial angle and large facets in the “acute zone” of the eye, where they need maximum resolution and sensitivity, and ommatidia with wider inter-ommatidial angle on the sides, top and bottom of the eye, where images move too fast to need a good resolution (Horridge and Duelli, 1979). Therefore, measuring inter-ommatidial angle in the pseudocone would be a good addition to our comprehensive analysis. In some of the

cleared eyes, the ommatidia lenses had been disattached from the ommatidia pigment cells, thus the original distance between the lens and the distal photoreceptor end could not be accurately measured. Improvements in the bleaching and clearing protocol that can reduce these artifacts and different incubation times could be applied to solve this problem and allow the measurement of this focal length.

5.4.4 Outlook

The next obvious step is to check if *oc* is indeed the evolved gene that causes eye size variation between *D. mauritiana* TAM16 and *D. simulans* YVF. In this direction, we want to take advantage of the CRISPR/Cas9 methodology (Barrangou et al., 2007; Jinek et al., 2012), which is currently being successfully applied in a large range of non-model organisms (Chen et al., 2014; Gilles et al., 2015; Sugano et al., 2014; Wang et al., 2013). We have used the model species *D. melanogaster* to analyze gene function using RNAi technique. However, to confirm the evolution in the sequence of the candidate gene we must use the species that present morphological differences. The aim is to generate flies with *D. simulans* YVF genetic background but containing the *oc* locus from *D. mauritiana* TAM16. If these flies have indeed bigger eyes than control *D. simulans* YVF flies we can confirm that changes in the *oc* locus are indeed responsible for the development of larger eyes in *D. mauritiana* TAM16.

To identify the regulatory region that has evolved to give rise to the different in *oc* expression between the two species, we could also apply ATAC-seq technology, which can detect genome-wide open chromatin regions (Buenrostro et al., 2013). We could use eye-antennal imaginal discs at 120h AEL, when we have seen that this gene is differentially expressed, and also at mid-pupation, when *rhodopsin* expression starts (Wernet et al., 2006). If a candidate regulatory region could be found, CRISPR/Cas9 could be used to replace only that sequence, therefore facilitating the procedure by reducing the length of the exchanged region.

Oc is the homolog of the CRX/OTX gene family of transcription factors in mammals, which has three known members in mouse, CRX, OTX1 and OTX2. It has been shown that these transcription factors, even though they are separated by a large evolutionary distance, are involved in almost the same biological processes, including their role in eye development (Freund et al., 1997; Furukawa et al., 1997), and even share most of their target genes (Ranade et al., 2008). This indicates a strong evolutionary conservation, especially at the coding sequence level. However, our findings, together with reports

indicating a large degree of natural variation in its *cis*-regulatory region of the *oc* gene in *D. melanogaster* populations (Goering et al., 2009), could favor the idea that its various *cis*-regulatory elements are the subject of strong positive selection, as it is the case for other homeobox genes or other components of the developmental “genetic toolkit” (e.g. Löhrl and Pick, 2005; McMahon et al., 2003; Ronshaugen et al., 2002; reviewed in Wagner and Lynch, 2008).

6 References

- Abu-Shaar, M., and Mann, R. S. (1998). Generation of multiple antagonistic domains along the proximodistal axis during *Drosophila* leg development. *Development* 125, 3821–30.
- Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–95.
- Alonso, C. R., and Wilkins, A. S. (2005). The molecular elements that underlie developmental evolution. *Nat. Rev. Genet.* 6, 709–15. doi:10.1038/nrg1676.
- Alsö, J. M., Tarchini, B., Cayouette, M., and Livesey, F. J. (2013). Ikaros promotes early-born neuronal fates in the cerebral cortex. *Proc. Natl. Acad. Sci. U. S. A.* 110, E716–25. doi:10.1073/pnas.1215707110.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–10. doi:10.1016/S0022-2836(05)80360-2.
- Alvarez-Ponce, D., Aguade, M., and Rozas, J. (2008). Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Res.* 19, 234–242. doi:10.1101/gr.084038.108.
- Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5350–4.
- Amore, G., and Casares, F. (2010). Size matters: The contribution of cell proliferation to the progression of the specification *Drosophila* eye gene regulatory network. *Dev. Biol.* 344, 569–577. doi:10.1016/j.ydbio.2010.06.015.
- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106. doi:10.1186/gb-2010-11-10-r106.
- Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq A Python framework to work with high-throughput sequencing data. *bioRxiv* 31, 002824. doi:10.1101/002824.
- Anderson, J., Salzer, C. L., and Kumar, J. P. (2006). Regulation of the retinal determination gene *dachshund* in the embryonic head and developing eye of *Drosophila*. *Dev. Biol.* 297, 536–549. doi:10.1016/j.ydbio.2006.05.004.
- Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437, 1149–52. doi:10.1038/nature04107.
- Apitz, H., and Salecker, I. (2014). A Challenge of Numbers and Diversity: Neurogenesis in the *Drosophila* Optic Lobe. *J. Neurogenet.*
- Arif, S., Hilbrant, M., Hopfen, C., Almudi, I., Nunes, M. D. S., Posnien, N., et al. (2013). Genetic and developmental analysis of differences in eye and face morphology between *Drosophila simulans* and *Drosophila mauritiana*. *Evol. Dev.* 15, 257–267. doi:10.1111/ede.12027.
- Ashburner, M. (1989). *Drosophila. A laboratory handbook*. Cold Spring Harbor Laboratory Press.
- Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D., and Hibberd, J. M. (2014). Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment

- of Trans-Factors in Two Independent Origins of C4 Photosynthesis. *PLoS Genet.* 10, e1004365. doi:10.1371/journal.pgen.1004365.
- Awasaki, T., Lai, S.-L., Ito, K., and Lee, T. (2008). Organization and Postembryonic Development of Glial Cells in the Adult Central Brain of *Drosophila*. *J. Neurosci.* 28, 13742–13753. doi:10.1523/JNEUROSCI.4844-08.2008.
- Baack, E. J., Sapir, Y., Chapman, M. A., Burke, J. M., and Rieseberg, L. H. (2008). Selection on domestication traits and quantitative trait loci in crop-wild sunflower hybrids. *Mol. Ecol.* 17, 666–77. doi:10.1111/j.1365-294X.2007.03596.x.
- Bainton, R. J., Tsai, L. T.-Y., Schwabe, T., DeSalvo, M., Gaul, U., and Heberlein, U. (2005). moody Encodes Two GPCRs that Regulate Cocaine Behaviors and Blood-Brain Barrier Permeability in *Drosophila*. *Cell* 123, 145–156. doi:10.1016/j.cell.2005.07.029.
- Baker, N. E. (2000). Notch signaling in the nervous system. Pieces still missing from the puzzle. *Bioessays* 22, 264–73. doi:10.1002/(SICI)1521-1878(200003)22:3<264::AID-BIES8>3.0.CO;2-M.
- Baonza, A., and Freeman, M. (2002). Control of *Drosophila* eye specification by Wingless signalling. *Development* 129, 5313–22.
- Baonza, A., Murawsky, C. M., Travers, A. A., and Freeman, M. (2002). Pointed and Tramtrack69 establish an EGFR-dependent transcriptional switch to regulate mitosis. *Nat. Cell Biol.* 4, 976–80. doi:10.1038/ncb887.
- Bardou, P., Mariette, J., Escudié, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, 1–7. doi:10.1186/1471-2105-15-293.
- Barolo, S., Carver, L. A., and Posakony, J. W. (2000). GFP and beta-galactosidase transformation vectors for promoter/enhancer analysis in *Drosophila*. *Biotechniques* 29, 726, 728, 730, 732.
- Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–12. doi:10.1126/science.1138140.
- Beckervordersandforth, R. M., Rickert, C., Altenhein, B., and Technau, G. M. (2008). Subtypes of glial cells in the *Drosophila* embryonic ventral nerve cord as related to lineage and gene expression. *Mech. Dev.* 125, 542–57. doi:10.1016/j.mod.2007.12.004.
- Bell, M. L., Earl, J. B., and Britt, S. G. (2007). Two types of *Drosophila* R7 photoreceptor cells are arranged randomly: a model for stochastic cell-fate determination. *J. Comp. Neurol.* 502, 75–85. doi:10.1002/cne.21298.
- Belting, H.-G., Shashikant, C. S., and Ruddle, F. H. (1998). Modification of expression and cis-regulation of Hoxc8 in the evolution of diverged axial morphology. *Proc. Natl. Acad. Sci.* 95, 2355–2360. doi:10.1073/pnas.95.5.2355.
- Bender, M., Turner, F. R., and Kaufman, T. C. (1987). A developmental genetic analysis of the gene Regulator of postbithorax in *Drosophila melanogaster*. *Dev Biol* 119, 418–432.
- Bock, I. R., and Wheeler, M. R. (1972). The *Drosophila melanogaster* species group. *Univ. Texas Publs Stud. Genet.* 7213, 1–102.
- Borevitz, J. O., and Chory, J. (2004). Genomics tools for QTL analysis and gene discovery. *Curr. Opin. Plant Biol.* 7, 132–6. doi:10.1016/j.pbi.2004.01.011.

- Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., et al. (2004). GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 20, 3710–5. doi:10.1093/bioinformatics/bth456.
- Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature* 478, 343–348. doi:10.1038/nature10532.
- Bray, N., Pimentel, H., Melsted, P., and Pachter, L. (2015). Near-optimal RNA-Seq quantification. *arXiv* 1505.02710. doi:arXiv:1505.02710.
- Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752–5. doi:10.1126/science.1069516.
- Brennan, C. A., and Moses, K. (2000). Determination of *Drosophila* photoreceptors: timing is everything. *Cell. Mol. Life Sci.* 57, 195–214.
- Britten, R., and Davidson, E. (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* 46, 111–138.
- Brown, N. L., Paddock, S. W., Sattler, C. a, Cronmiller, C., Thomas, B. J., and Carroll, S. B. (1996). daughterless is required for *Drosophila* photoreceptor cell determination, eye morphogenesis, and cell cycle progression. *Dev. Biol.* 179, 65–78. doi:10.1006/dbio.1996.0241.
- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–8. doi:10.1038/nmeth.2688.
- Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11, 94. doi:10.1186/1471-2105-11-94.
- Burla, H. (1951). Systematik, Verbreitung und Oekologie der *Drosophila*-Artender Schweiz. *Rev. suisse Zool.* 58, 23–175.
- Burnet, B., Conolly, K., and Beck, J. (1967). Phenogenetic studies on visual acuity in *Drosophila melanogaster*. *J. Insect Physiol.* 14, 855–860.
- Busby, M. A., Gray, J. M., Costa, A. M., Stewart, C., Stromberg, M. P., Barnett, D., et al. (2011). Expression divergence measured by transcriptome sequencing of four yeast species. *BMC Genomics* 12, 635. doi:10.1186/1471-2164-12-635.
- Bustin, S. A. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J. Mol. Endocrinol.* 25, 169–93.
- Cagan, R. L., and Ready, D. F. (1989). The emergence of order in the *Drosophila* pupal retina. *Dev. Biol.* 136, 346–62.
- Campbell, G., and Tomlinson, A. (1998). The roles of the homeobox genes *aristaless* and *Distal-less* in patterning the legs and wings of *Drosophila*. *Development* 125, 4483–93.
- Cao, W., Song, H.-J., Gangi, T., Kelkar, A., Antani, I., Garza, D., et al. (2008). Identification of novel genes that modify phenotypes induced by Alzheimer's beta-amyloid overexpression in *Drosophila*. *Genetics* 178, 1457–71. doi:10.1534/genetics.107.078394.

- Carlson, M. R. J., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., and Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* 7, 40. doi:10.1186/1471-2164-7-40.
- Carlson, S. D., Juang, J. L., Hilgers, S. L., and Garment, M. B. (2000). Blood barriers of the insect. *Annu. Rev. Entomol.* 45, 151–74. doi:10.1146/annurev.ento.45.1.151.
- Carroll, S. B. (1995). Homeotic genes and the evolution of arthropods and chordates. *Nature* 376, 479–85. doi:10.1038/376479a0.
- Carroll, S. B. (2001). *From DNA to diversity*. Blackwell Science.
- Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134, 25–36. doi:10.1016/j.cell.2008.06.030.
- Casares, F., and Mann, R. S. (1998). Control of antennal versus leg development in *Drosophila*. *Nature* 392, 723–6. doi:10.1038/33706.
- Celniker, S. E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., et al. (2009). Unlocking the secrets of the genome. *Nature* 459, 927–30. doi:10.1038/459927a.
- Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., et al. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science* 327, 302–5. doi:10.1126/science.1182213.
- Chapman, R. (1998). *The Insects: Structure and Function*. 4th ed. Cambridge University Press.
- Chen, L., Tang, L., Xiang, H., Jin, L., Li, Q., Dong, Y., et al. (2014). Advances in genome editing technology and its promising application in evolutionary and ecological studies. *Gigascience* 3, 24. doi:10.1186/2047-217X-3-24.
- Cheverud, J., and Routman, E. (1993). Quantitative trait loci - individual gene effects on quantitative characters. *J. Evol. Biol.* 6, 463–480.
- Chhangawala, S., Rudy, G., Mason, C. E., and Rosenfeld, J. A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biol.* 16, 131. doi:10.1186/s13059-015-0697-y.
- Ching, T., Huang, S., and Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* 20, 1684–96. doi:10.1261/rna.046011.114.
- Cho, K. O., and Choi, K. W. (1998). Fringe is essential for mirror symmetry and morphogenesis in the *Drosophila* eye. *Nature* 396, 272–6. doi:10.1038/24394.
- Choi, K. W., and Benzer, S. (1994). Migration of glia along photoreceptor axons in the developing *Drosophila* eye. *Neuron* 12, 423–431. doi:10.1016/0896-6273(94)90282-8.
- Chou, W. H., Huber, A., Bentrop, J., Schulz, S., Schwab, K., Chadwell, L. V., et al. (1999). Patterning of the R7 and R8 photoreceptor cells of *Drosophila*: evidence for induced and default cell-fate specification. *Development* 126, 607–16.
- Chu, C., Fang, Z., Hua, X., Yang, Y., Chen, E., Cowley, A. W., et al. (2015). deGPS is a powerful tool for detecting differential expression in RNA-sequencing studies. *BMC Genomics* 16, 455. doi:10.1186/s12864-015-1676-0.
- Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* 2, 2366–2382. doi:10.1038/nprot.2007.324.

- Coolon, J. D., Mcmanus, C. J., Stevenson, K. R., Coolon, J. D., Mcmanus, C. J., Stevenson, K. R., et al. (2014). Tempo and mode of regulatory evolution in *Drosophila*. 797–808. doi:10.1101/gr.163014.113.
- Coolon, J. D., and Wittkopp, P. J. (2013). cis- and trans- regulation in *Drosophila* interspecific hybrids. In: *Polyloid and Hybrid Genomics.*, eds. Z. J. Chen and J. A. Birchler Oxford, UK: John Wiley & Sons, Inc. doi:10.1002/9781118552872.
- Cowles, C. R., Hirschhorn, J. N., Altshuler, D., and Lander, E. S. (2002). Detection of regulatory variation in mouse genes. *Nat. Genet.* 32, 432–437. doi:10.1038/ng992.
- Curtiss, J., and Mlodzik, M. (2000). Morphogenetic furrow initiation and progression during eye development in *Drosophila*: the roles of decapentaplegic, hedgehog and eyes absent. *Development* 127, 1325–36.
- Dai, P., Akimaru, H., Tanaka, Y., Hou, D. X., Yasukawa, T., Kanei-Ishii, C., et al. (1996). CBP as a transcriptional coactivator of c-Myb. *Genes Dev.* 10, 528–40.
- Davidson, E. H. (2001). *Genomic regulatory systems: development and evolution*. Academic, San Diego.
- Davidson, E. H. (2006). Gene Regulatory Networks and the Evolution of Animal Body Plans. *Science* 311, 796–800. doi:10.1126/science.1113832.
- Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., et al. (2002). A genomic regulatory network for development. *Science* 295, 1669–78. doi:10.1126/science.1069883.
- Davila-Velderrain, J., Servin-Marquez, A., and Alvarez-Buylla, E. R. (2014). Molecular evolution constraints in the floral organ specification gene regulatory network module across 18 angiosperm genomes. *Mol. Biol. Evol.* 31, 560–73. doi:10.1093/molbev/mst223.
- Dearborn, R., and Kunes, S. (2004). An axon scaffold induced by retinal axons directs glia to destinations in the *Drosophila* optic lobe. *Development* 131, 2291–303. doi:10.1242/dev.01111.
- DeLuca, S. Z., and O’Farrell, P. H. (2012). Barriers to male transmission of mitochondrial DNA in sperm development. *Dev. Cell* 22, 660–8. doi:10.1016/j.devcel.2011.12.021.
- DeSalvo, M. K., Hindle, S. J., Rusan, Z. M., Orng, S., Eddison, M., Halliwill, K., et al. (2014). The *Drosophila* surface glia transcriptome: evolutionary conserved blood-brain barrier processes. *Front. Neurosci.* 8, 1–22. doi:10.3389/fnins.2014.00346.
- DeSalvo, M. K., Mayer, N., Mayer, F., and Bainton, R. J. (2011). Physiologic and anatomic characterization of the brain surface glia barrier of *Drosophila*. *Glia* 59, 1322–40. doi:10.1002/glia.21147.
- Dey, B. K., Zhao, X.-L., Popo-Ola, E., and Campos, A. R. (2009). Mutual regulation of the *Drosophila* disconnected (disco) and Distal-less (Dll) genes contributes to proximal-distal patterning of antenna and leg. *Cell Tissue Res.* 338, 227–40. doi:10.1007/s00441-009-0865-z.
- Dietzl, G., Chen, D., Schnorrer, F., Su, K.-C., Barinova, Y., Fellner, M., et al. (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* 448, 151–6. doi:10.1038/nature05954.
- Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2012). A comprehensive evaluation of normalization methods for Illumina high-

- throughput RNA sequencing data analysis. *Brief. Bioinform.*, bbs046. doi:10.1093/bib/bbs046.
- Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., et al. (2007). A genome-wide association study of global gene expression. *Nat. Genet.* 39, 1202–7. doi:10.1038/ng2109.
- Domínguez, M., and Casares, F. (2005). Organ specification-growth control connection: new in-sights from the *Drosophila* eye-antennal disc. *Dev. Dyn.* 232, 673–84. doi:10.1002/dvdy.20311.
- Domínguez, M., and Hafen, E. (1997). Hedgehog directly controls initiation and propagation of retinal differentiation in the *Drosophila* eye. *Genes Dev.* 11, 3254–64.
- Dong, P. D., Chu, J., and Panganiban, G. (2000). Coexpression of the homeobox genes *Distal-less* and *homothorax* determines *Drosophila* antennal identity. *Development* 127, 209–16.
- Earl, J. B., and Britt, S. G. (2006). Expression of *Drosophila* rhodopsins during photoreceptor cell differentiation: insights into R7 and R8 cell subtype commitment. *Gene Expr. Patterns* 6, 687–94. doi:10.1016/j.modgep.2006.01.003.
- Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–10.
- Edwards, J. S., Swales, L. S., and Bate, M. (1993). The differentiation between neuroglia and connective tissue sheath in insect ganglia revisited: the neural lamella and perineurial sheath cells are absent in a mesodermless mutant of *Drosophila*. *J. Comp. Neurol.* 333, 301–8. doi:10.1002/cne.903330214.
- Edwards, T. N., Nuschke, A. C., Nern, A., and Meinertzhagen, I. A. (2012). Organization and metamorphosis of glia in the *Drosophila* visual system. *J. Comp. Neurol.* 520, 2067–85. doi:10.1002/cne.23071.
- Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* 29, 51–63. doi:10.1016/j.tree.2013.09.008.
- Elliott, J., Jolicoeur, C., Ramamurthy, V., and Cayouette, M. (2008). Ikaros Confers Early Temporal Competence to Mouse Retinal Progenitor Cells. *Neuron* 60, 26–39. doi:10.1016/j.neuron.2008.08.008.
- Erickson, D. L., Fenster, C. B., Stenøien, H. K., and Price, D. (2004). Quantitative trait locus analyses and the study of evolutionary process. *Mol. Ecol.* 13, 2505–22. doi:10.1111/j.1365-294X.2004.02254.x.
- Erives, A., and Levine, M. (2004). Coordinate enhancers share common organizational features in the *Drosophila* genome. *Proc. Natl. Acad. Sci. U. S. A.* 101, 3851–6. doi:10.1073/pnas.0400611101.
- Evans, J. D., Brown, S. J., Hackett, K. J. J., Robinson, G., Richards, S., Lawson, D., et al. (2013). The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Hered.* 104, 595–600. doi:10.1093/jhered/est050.
- Falconer, D., and Mackay, T. (1995). *Introduction to quantitative genetics*. 4th editio. Addison-Wesley Longman, Harlow, UK.
- Fan, J.-B., Yeakley, J. M., Bibikova, M., Chudin, E., Wickham, E., Chen, J., et al. (2004). A versatile assay for high-throughput gene expression profiling on universal array matrices. *Genome Res.* 14, 878–85. doi:10.1101/gr.2167504.

- Fichelson, P., Brigui, A., and Pichaud, F. (2012). Orthodenticle and Kruppel homolog 1 regulate *Drosophila* photoreceptor maturation. *Proc. Natl. Acad. Sci. U. S. A.* 109, 7893–8. doi:10.1073/pnas.1120276109.
- Filteau, M., Pavey, S. A., St-Cyr, J., and Bernatchez, L. (2013). Gene Coexpression Networks Reveal Key Drivers of Phenotypic Divergence in Lake Whitefish. *Mol. Biol. Evol.* 30, 1384–1396. doi:10.1093/molbev/mst053.
- Finkelstein, R., and Boncinelli, E. (1994). From fly head to mammalian forebrain: the story of otd and Otx. *Trends Genet.* 10, 310–5.
- Finkelstein, R., Smouse, D., Capaci, T. M., Spradling, a C., and Perrimon, N. (1990). The orthodenticle gene encodes a novel homeo domain protein involved in the development of the *Drosophila* nervous system and ocellar visual structures. *Genes Dev.* 4, 1516–1527. doi:10.1101/gad.4.9.1516.
- Fonseca, N. a., Morales-Hojas, R., Reis, M., Rocha, H., Vieira, C. P., Nolte, V., et al. (2013). *Drosophila americana* as a Model Species for Comparative Studies on the Molecular Basis of Phenotypic Variation. *Genome Biol. Evol.* 5, 661–679. doi:10.1093/gbe/evt037.
- Fontanillas, P., Landry, C. R., Wittkopp, P. J., Russ, C., Gruber, J. D., Nusbaum, C., et al. (2010). Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Mol. Ecol.* 19, 212–227. doi:10.1111/j.1365-294X.2010.04472.x.
- Fortini, M. E., and Rubin, G. M. (1990). Analysis of cis-acting requirements of the Rh3 and Rh4 genes reveals a bipartite organization to rhodopsin promoters in *Drosophila melanogaster*. *Genes Dev.* 4, 444–63.
- Frankfort, B. J., and Mardon, G. (2002). R8 development in the *Drosophila* eye: a paradigm for neural selection and differentiation. *Development* 129, 1295–1306.
- Frary, A., Nesbitt, T. C., Grandillo, S., Knaap, E., Cong, B., Liu, J., et al. (2000). fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. *Science* 289, 85–8.
- Freeman, M. (1994). The spitz gene is required for photoreceptor determination in the *Drosophila* eye where it interacts with the EGF receptor. *Mech. Dev.* 48, 25–33.
- Freeman, M. (1996). Reiterative Use of the EGF Receptor Triggers Differentiation of All Cell Types in the *Drosophila* Eye. *Cell* 87, 651–660. doi:10.1016/S0092-8674(00)81385-9.
- Freeman, M. (1997). Cell determination strategies in the *Drosophila* eye. *Development* 124, 261–70.
- Freund, C. L., Gregory-Evans, C. Y., Furukawa, T., Papaioannou, M., Looser, J., Ploder, L., et al. (1997). Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell* 91, 543–53.
- Fristrom, D., and Fristrom, J. W. (1975). The mechanism of evagination of imaginal discs of *Drosophila melanogaster*. 1. General considerations. *Dev. Biol.* 43, 1–23.
- Fristrom, D., and Fristrom, J. W. (1993). *The metamorphic development of the adult epidermis. In: The development of Drosophila melanogaster.* Cold Spring Harbor Laboratory Press, New York.
- Fromental-Ramain, C., Vanolst, L., Delaporte, C., and Ramain, P. (2008). pannier encodes two structurally related isoforms that are differentially expressed during *Drosophila*

- development and display distinct functions during thorax patterning. *Mech. Dev.* 125, 43–57. doi:10.1016/j.mod.2007.10.008.
- Furukawa, T., Morrow, E. M., and Cepko, C. L. (1997). Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell* 91, 531–41.
- Galant, R., Walsh, C. M., and Carroll, S. B. (2002). Hox repression of a target gene: extradenticle-independent, additive action through multiple monomer binding sites. *Development* 129, 3115–26.
- Gall, J. G., and Pardue, M. L. (1969). Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proc. Natl. Acad. Sci. U. S. A.* 63, 378–83.
- Gao, H., Chen, X., Du, X., Guan, B., Liu, Y., and Zhang, H. (2011). EGF enhances the migration of cancer cells by up-regulation of TRPM7. *Cell Calcium* 50, 559–68. doi:10.1016/j.ceca.2011.09.003.
- Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477. doi:10.1038/nmeth.1613.
- Gehring, W. J. (2002). The genetic control of eye development and its implications for the evolution of the various eye-types. *Int. J. Dev. Biol.* 46, 65–73.
- Gerstein, M. B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J. B., et al. (2014). Comparative analysis of the transcriptome across distant species. *Nature* 512, 445–448. doi:10.1038/nature13424.
- Gibson, G., and Weir, B. (2005). The quantitative genetics of transcription. *Trends Genet.* 21, 616–23. doi:10.1016/j.tig.2005.08.010.
- Gidaszewski, N. A., Baylac, M., and Klingenberg, C. P. (2009). Evolution of sexual dimorphism of wing shape in the *Drosophila melanogaster* subgroup. *BMC Evol. Biol.* 9, 110. doi:10.1186/1471-2148-9-110.
- Gilles, A. F., Schinko, J. B., and Averof, M. (2015). Efficient CRISPR-mediated gene targeting and transgene replacement in the beetle *Tribolium castaneum*. *Development* 142, 2832–9. doi:10.1242/dev.125054.
- Girke, T. (2015). systemPipeR: NGS workflow and report generation environment. R package version 1.4.8, <https://github.com/tgirke/systemPipeR>.
- Goering, L. M., Hunt, P. K., Heighington, C., Busick, C., Pennings, P., Hermisson, J., et al. (2009). Association of orthodenticle with natural variation for early embryonic patterning in *Drosophila melanogaster*. *J. Exp. Zool. Part B Mol. Dev. Evol.* 312B, 841–854. doi:10.1002/jez.b.21299.
- Goncalves, A., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., et al. (2012). Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res.* 22, 2376–2384. doi:10.1101/gr.142281.112.
- González, E., and Joly, S. (2013). Impact of RNA-seq attributes on false positive rates in differential expression analysis of de novo assembled transcriptomes. *BMC Res. Notes* 6, 503. doi:10.1186/1756-0500-6-503.
- Gould, S. (1977). *Ontogeny and Phylogeny*. Belknap, Cambridge, MA.
- Graze, R. M., McIntyre, L. M., Main, B. J., Wayne, M. L., and Nuzhdin, S. V. (2009). Regulatory Divergence in *Drosophila melanogaster* and *D. simulans*, a Genomewide

- Analysis of Allele-Specific Expression. *Genetics* 183, 547–561.
doi:10.1534/genetics.109.105957.
- Graze, R. M., Novelo, L. L., Amin, V., Fear, J. M., Casella, G., Nuzhdin, S. V., et al. (2012). Allelic Imbalance in *Drosophila* Hybrid Heads: Exons, Isoforms, and Evolution. *Mol. Biol. Evol.* 29, 1521–1532. doi:10.1093/molbev/msr318.
- Grens, A., Mason, E., Marsh, J. L., and Bode, H. R. (1995). Evolutionary conservation of a cell fate specification gene: the Hydra achaete-scute homolog has proneural activity in *Drosophila*. *Development* 121, 4027–35.
- Griffin, C., Kleinjan, D. A., Doe, B., and van Heyningen, V. (2002). New 3' elements control Pax6 expression in the developing pretectum, neural retina and olfactory region. *Mech. Dev.* 112, 89–100.
- Grimaldi, D. A. (1987). Phylogenetics and taxonomy of Zygothrica (Diptera: Drosophilidae). *Bull. Am. Mus. nat. Hist.* 186, 103–268.
- Grosskortenhause, R., Pearson, B. J., Marusich, A., and Doe, C. Q. (2005). Regulation of temporal identity transitions in *Drosophila* neuroblasts. *Dev. Cell* 8, 193–202. doi:10.1016/j.devcel.2004.11.019.
- Gunthorpe, D., Beatty, K. E., and Taylor, M. V (1999). Different levels, but not different isoforms, of the *Drosophila* transcription factor DMEF2 affect distinct aspects of muscle differentiation. *Dev. Biol.* 215, 130–45. doi:10.1006/dbio.1999.9449.
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–512. doi:10.1038/nprot.2013.084.
- Halder, G., Callaerts, P., and Gehring, W. J. (1995). Induction of ectopic eyes by targeted expression of the eyeless gene in *Drosophila*. *Science* 267, 1788–92.
- Hansen, K. D., Wu, Z., Irizarry, R. A., and Leek, J. T. (2011). Sequencing technology does not eliminate biological variability. *Nat. Biotechnol.* 29, 572–3. doi:10.1038/nbt.1910.
- Hardy, R. C. (1985). Functional organization of the fly retina. *Prog. Sens. Physiol.* 5, 1–79.
- Haug, J. T., Haug, C., Kutschera, V., Mayer, G., Maas, A., Liebau, S., et al. (2011). Autofluorescence imaging, an excellent tool for comparative morphology. *J. Microsc.* 244, 259–72. doi:10.1111/j.1365-2818.2011.03534.x.
- Hausler, D., O'Brien, S. J., Ryder, O. a., Keith Barker, F., Clamp, M., Crawford, A. J., et al. (2009). Genome 10K: A proposal to obtain whole-genome sequence for 10000 vertebrate species. *J. Hered.* 100, 659–674. doi:10.1093/jhered/esp086.
- Hayes, B., and Goddard, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33, 209–229. doi:10.1186/1297-9686-33-3-209.
- Haynie, J. L., and Bryant, P. J. (1986). Development of the eye-antenna imaginal disc and morphogenesis of the adult head in *Drosophila melanogaster*. *J. Exp. Zool.* 237, 293–308. doi:10.1002/jez.1402370302.
- Hazelett, D. J., Bourouis, M., Walldorf, U., and Treisman, J. E. (1998). decapentaplegic and wingless are regulated by eyes absent and eyegone and interact to direct the pattern of retinal differentiation in the eye disc. *Development* 125, 3741–51.

- Heanue, T. A., Reshef, R., Davis, R. J., Mardon, G., Oliver, G., Tomarev, S., et al. (1999). Synergistic regulation of vertebrate muscle development by Dach2, Eya2, and Six1, homologs of genes required for *Drosophila* eye formation. *Genes Dev.* 13, 3231–43.
- Hecht, S., and Wolf, E. (1929). The visual acuity of the honey bee. *J. Gen. Physiol.* 12, 727–60.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–89. doi:10.1016/j.molcel.2010.05.004.
- Herrmann, C., Van de Sande, B., Potier, D., and Aerts, S. (2012). i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res.* 40, e114–e114. doi:10.1093/nar/gks543.
- Hewitt, Z., Forsyth, N. R., Waterfall, M., Wojtacha, D., Thomson, A. J., and McWhir, J. (2006). Fluorescence-activated single cell sorting of human embryonic stem cells. *Cloning Stem Cells* 8, 225–34. doi:10.1089/clo.2006.8.225.
- Hilbrant, M., Almudi, I., Leite, D. J., Kuncheria, L., Posnien, N., Nunes, M. D. S., et al. (2014). Sexual dimorphism and natural variation within and among species in the *Drosophila* retinal mosaic. *BMC Evol. Biol.* 14, 240. doi:10.1186/s12862-014-0240-x.
- Von Hilchen, C. M., Bustos, A. E., Giangrande, A., Technau, G. M., and Altenhein, B. (2013). Predetermined embryonic glial cells form the distinct glial sheaths of the *Drosophila* peripheral nervous system. *Development* 140, 3657–68. doi:10.1242/dev.093245.
- Hilscher, J., Schlötterer, C., and Hauser, M.-T. (2009). A single amino acid replacement in ETC2 shapes trichome patterning in natural *Arabidopsis* populations. *Curr. Biol.* 19, 1747–51. doi:10.1016/j.cub.2009.08.057.
- Hirono, K., Margolis, J. S., Posakony, J. W., and Doe, C. Q. (2012). Identification of hunchback cis-regulatory DNA conferring temporal expression in neuroblasts and neurons. *Gene Expr. Patterns* 12, 11–17. doi:10.1016/j.gep.2011.10.001.
- Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2015). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32, 767–769. doi:10.1093/bioinformatics/btv661.
- Homem, C. C. F., and Knoblich, J. A. (2012). *Drosophila* neuroblasts: a model for stem cell biology. *Development* 139, 4297–4310. doi:10.1242/dev.080515.
- Hornett, E. A., and Wheat, C. W. (2012). Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC Genomics* 13, 361. doi:10.1186/1471-2164-13-361.
- Horridge, G. A. (1978). The Separation of Visual Axes in Apposition Compound Eyes. *Philos. Trans. R. Soc. B Biol. Sci.* 285, 1–59. doi:10.1098/rstb.1978.0093.
- Horridge, G. A., and Duelli, P. (1979). Anatomy of the Regional Differences in the Eye of the Mantis *Ciulfina*. *J. Exp. Biol.* 80, 165–190.
- Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., et al. (2007). Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* 316, 1625–8. doi:10.1126/science.1139816.

- Hu, T. T., Eisen, M. B., Thornton, K. R., and Andolfatto, P. (2013). A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Res.* 23, 89–98. doi:10.1101/gr.141689.112.
- Hummel, T., Attix, S., Gunning, D., Zipursky, S. L., and Angeles, L. (2002). Temporal Control of Glial Cell Migration in the. 33, 193–203.
- Iadecola, C., and Nedergaard, M. (2007). Glial regulation of the cerebral microvasculature. *Nat. Neurosci.* 10, 1369–76. doi:10.1038/nn2003.
- Isshiki, T., Pearson, B., Holbrook, S., and Doe, C. Q. (2001). *Drosophila* Neuroblasts Sequentially Express Transcription Factors which Specify the Temporal Identity of Their Neuronal Progeny. *Cell* 106, 511–521. doi:10.1016/S0092-8674(01)00465-2.
- Ito, K., Urban, J., and Technau, G. M. (1995). Distribution, classification, and development of *Drosophila* glial cells in the late embryonic and early larval ventral nerve cord. *Roux's Arch. Dev. Biol.* 204, 284–307. doi:10.1007/BF02179499.
- Jenny, A. (2011). Preparation of Adult *Drosophila* Eyes for Thin Sectioning and Microscopic Analysis. *J. Vis. Exp.*, 1–5. doi:10.3791/2959.
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–21. doi:10.1126/science.1225829.
- Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497–502. doi:10.1126/science.1141319.
- Johnson, N. A. (2007). The Micro-evolution of development. *Genetica* 129, 1–5. doi:10.1007/s10709-006-0028-z.
- Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E. H., Birney, E., et al. (2012). A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. *Cell* 148, 473–486. doi:10.1016/j.cell.2012.01.030.
- Jurgens, G., and Hartenstein, V. (1993). *The terminal regions of the body pattern*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Juven-Gershon, T., Hsu, J.-Y., and Kadonaga, J. T. (2008). Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes Dev.* 22, 2823–30. doi:10.1101/gad.1698108.
- Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–9. doi:10.1093/bioinformatics/bts199.
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006. doi:10.1101/gr.229102.
- Kenyon, K. L., Ranade, S. S., Curtiss, J., Mlodzik, M., and Pignoni, F. (2003). Coordinating proliferation and tissue specification to promote regional identity in the *Drosophila* head. *Dev. Cell* 5, 403–14.

- Khaitovich, P., Enard, W., Lachmann, M., and Pääbo, S. (2006). Evolution of primate gene expression. *Nat. Rev. Genet.* 7, 693–702. doi:10.1038/nrg1940.
- Kim, J., Kerr, J. Q., and Min, G.-S. (2000). Molecular heterochrony in the early development of *Drosophila*. *Proc. Natl. Acad. Sci.* 97, 212–216. doi:10.1073/pnas.97.1.212.
- King, M. C., and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science* 188, 107–116.
- Klaus, A. V., Kulasekera, V. L., and Schawaroch, V. (2003). Three-dimensional visualization of insect morphology using confocal laser scanning microscopy. *J. Microsc.* 212, 107–116.
- Klein, T. (2008). Immunolabeling of imaginal discs. *Methods Mol. Biol.* 420, 253–263. doi:10.1007/978-1-59745-583-1_15.
- Knight, J. C. (2004). Allele-specific gene expression uncovered. *Trends Genet.* 20, 113–116. doi:10.1016/j.tig.2004.01.001.
- Koepfli, K.-P., Paten, B., and O'Brien, S. J. (2015). The Genome 10K Project: A Way Forward. *Annu. Rev. Anim. Biosci.* 3, 57–111. doi:10.1146/annurev-animal-090414-014900.
- Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., et al. (2011). PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS One* 6, e15925. doi:10.1371/journal.pone.0015925.
- Kopp, A., and McIntyre, L. M. (2012). Transcriptional network structure has little effect on the rate of regulatory evolution in yeast. *Mol. Biol. Evol.* 29, 1899–905. doi:10.1093/molbev/msq283.
- Kosman, D., Small, S., and Reinitz, J. (1998). Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. *Dev. Genes Evol.* 208, 290–294. doi:10.1007/s004270050184.
- Kumar, J. P. (2004). CREB Binding Protein Functions During Successive Stages of Eye Development in *Drosophila*. *Genetics* 168, 877–893. doi:10.1534/genetics.104.029850.
- Kumar, J. P. (2009). The Molecular Circuitry Governing Retinal Determination. *Biochim. Biophys. Acta* 1789, 306–314. doi:10.1016/j.bbagr.2008.10.001.
- Kumar, J. P., and Moses, K. (2000). Cell fate specification in the *Drosophila* retina. *Results Probl. Cell Differ.* 31, 93–114.
- Kumar, J. P., and Moses, K. (2001). Eye specification in *Drosophila*: perspectives and implications. *Semin. Cell Dev. Biol.* 12, 469–74. doi:10.1006/scdb.2001.0270.
- Kwok, R. P., Lundblad, J. R., Chrivia, J. C., Richards, J. P., Bächinger, H. P., Brennan, R. G., et al. (1994). Nuclear protein CBP is a coactivator for the transcription factor CREB. *Nature* 370, 223–6. doi:10.1038/370223a0.
- Lachaise, D., Cariou, M. L., David, J. R., Lemeunier, F., Tsacas, L., and Ashburner, M. (1988). Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evol. Biol.* 22, 159–225.
- Land, M. F., Gibson, G., Horwood, J., and Zeil, J. (1999). Fundamental differences in the optical structure of the eyes of nocturnal and diurnal mosquitoes. *J. Comp. Physiol. A Sensory, Neural, Behav. Physiol.* 185, 91–103. doi:10.1007/s003590050369.

- Landry, C. R., Wittkopp, P. J., Taubes, C. H., Ranz, J. M., Clark, A. G., and Hartl, D. L. (2005). Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. *Genetics* 171, 1813–22. doi:10.1534/genetics.105.047449.
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25. doi:10.1186/gb-2009-10-3-r25.
- Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15, R29. doi:10.1186/gb-2014-15-2-r29.
- Lee, H.-H., and Frasch, M. (2005). Nuclear integration of positive Dpp signals, antagonistic Wg inputs and mesodermal competence factors during *Drosophila* visceral mesoderm induction. *Development* 132, 1429–42. doi:10.1242/dev.01687.
- Lehmann, R., and Nüsslein-Volhard, C. (1987). hunchback, a gene required for segmentation of an anterior and posterior region of the *Drosophila* embryo. *Dev. Biol.* 119, 402–17.
- Lemon, B. (2000). Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* 14, 2551–2569. doi:10.1101/gad.831000.
- Leonard, D. S., Bowman, V. D., Ready, D. F., and Pak, W. L. (1992). Degeneration of photoreceptors in rhodopsin mutants of *Drosophila*. *J. Neurobiol.* 23, 605–26. doi:10.1002/neu.480230602.
- Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323. doi:10.1186/1471-2105-12-323.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.
- Li, T., and Bender, M. (2000). A conditional rescue system reveals essential functions for the ecdysone receptor (EcR) gene during molting and metamorphosis in *Drosophila*. *Development* 127, 2897–905.
- Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., et al. (2008). Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* 6, e27. doi:10.1371/journal.pbio.0060027.
- Lilly, B., Galewsky, S., Firulli, A. B., Schulz, R. A., and Olson, E. N. (1994). D-MEF2: a MADS box transcription factor expressed in differentiating mesoderm and muscle cell lineages during *Drosophila* embryogenesis. *Proc. Natl. Acad. Sci. U. S. A.* 91, 5662–6.
- Lim, J., Jafar-Nejad, H., Hsu, Y.-C., and Choi, K.-W. (2008). Novel function of the class I bHLH protein Daughterless in the negative regulation of proneural gene expression in the *Drosophila* eye. *EMBO Rep.* 9, 1128–33. doi:10.1038/embor.2008.166.

- Lin, Y., Golovkina, K., Chen, Z.-X., Lee, H. N., Negron, Y. L. S., Sultana, H., et al. (2016). Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics* 17, 28. doi:10.1186/s12864-015-2353-z.
- Liu, B. (1998). *Genomics, Linkage Mapping and QTL Analysis*. CRC Press, Boca Raton, FL.
- Liu, Y., Zhou, J., and White, K. P. (2014). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* 30, 301–4. doi:10.1093/bioinformatics/btt688.
- Loehlin, D. W., and Werren, J. H. (2012). Evolution of shape by multiple regulatory changes to a growth gene. *Science* 335, 943–7. doi:10.1126/science.1215193.
- Löhr, U., and Pick, L. (2005). Cofactor-interaction motifs and the cooption of a homeotic Hox protein into the segmentation pathway of *Drosophila melanogaster*. *Curr. Biol.* 15, 643–9. doi:10.1016/j.cub.2005.02.048.
- Lopes, C. S., and Casares, F. (2015). Eye selector logic for a coordinated cell cycle exit. *PLoS Genet.* 11, e1004981. doi:10.1371/journal.pgen.1004981.
- Love, M. I., Anders, S., and Huber, W. (2014a). Differential analysis of count data - the DESeq2 package. 1–41. doi:10.1101/002832.
- Love, M. I., Huber, W., and Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8.
- Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative traits*. Sinauer Associates, Sunderland, MA.
- Ma, C., and Moses, K. (1995). Wingless and patched are negative regulators of the morphogenetic furrow and can affect tissue polarity in the developing *Drosophila* compound eye. *Development* 121, 2279–89.
- Macmanes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Front. Genet.* 5, 13. doi:10.3389/fgene.2014.00013.
- Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448–9. doi:10.1093/bioinformatics/bti551.
- Malartre, M. (2016). Regulatory mechanisms of EGFR signalling during *Drosophila* eye development. *Cell. Mol. Life Sci.* doi:10.1007/s00018-016-2153-x.
- Malicki, J., Schughart, K., and McGinnis, W. (1990). Mouse Hox-2.2 specifies thoracic segmental identity in *Drosophila* embryos and larvae. *Cell* 63, 961–7.
- Malone, J. H., and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* 9, 34. doi:10.1186/1741-7007-9-34.
- Manceau, M., Domingues, V. S., Mallarino, R., and Hoekstra, H. E. (2011). The developmental role of Agouti in color pattern evolution. *Science* 331, 1062–5. doi:10.1126/science.1200684.
- Mao, Y., and Freeman, M. (2009). Fasciclin 2, the *Drosophila* orthologue of neural cell-adhesion molecule, inhibits EGF receptor signalling. *Development* 136, 473–81. doi:10.1242/dev.026054.

- Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi:10.1038/nmeth.2016.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bembien, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–80. doi:10.1038/nature03959.
- Marrone, A. K., Kucherenko, M. M., Rishko, V. M., and Shcherbata, H. R. (2011). New dystrophin/dystroglycan interactors control neuron behavior in *Drosophila* eye. *BMC Neurosci.* 12, 93. doi:10.1186/1471-2202-12-93.
- Maurel-Zaffran, C., and Treisman, J. E. (2000). pannier acts upstream of wingless to direct dorsal eye disc development in *Drosophila*. *Development* 127, 1007–1016.
- Mazzoni, E. O., Celik, A., Wernet, M. F., Vasilias, D., Johnston, R. J., Cook, T. A., et al. (2008). Iroquois Complex Genes Induce Co-Expression of rhodopsins in *Drosophila*. *PLoS Biol.* 6, e97. doi:10.1371/journal.pbio.0060097.
- McDermott, S. R., and Kliman, R. M. (2008). Estimation of isolation times of the island species in the *Drosophila simulans* complex from multilocus DNA sequence data. *PLoS One* 3, e2442. doi:10.1371/journal.pone.0002442.
- McDonald, E. C., Xie, B., Workman, M., Charlton-Perkins, M., Terrell, D. a., Reischl, J., et al. (2010). Separable transcriptional regulatory domains within Otd control photoreceptor terminal differentiation events. *Dev. Biol.* 347, 122–132. doi:10.1016/j.ydbio.2010.08.016.
- McGinnis, N., Kuziora, M. A., and McGinnis, W. (1990). Human Hox-4.2 and *Drosophila* deformed encode similar regulatory specificities in *Drosophila* embryos and larvae. *Cell* 63, 969–76.
- McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A., and Gehring, W. J. (1984). A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* 37, 403–8.
- McGurk, L., Morrison, H., Keegan, L. P., Sharpe, J., and O’Connell, M. A. (2007). Three-dimensional imaging of *Drosophila melanogaster*. *PLoS One* 2, e834. doi:10.1371/journal.pone.0000834.
- McMahon, A. P., Ingham, P. W., and Tabin, C. J. (2003). Developmental roles and clinical significance of hedgehog signaling. *Curr. Top. Dev. Biol.* 53, 1–114.
- McManus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., and Wittkopp, P. J. (2010). Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res.* 20, 816–825. doi:10.1101/gr.102491.109.
- McManus, K. J., and Hendzel, M. J. (2001). CBP, a transcriptional coactivator and acetyltransferase. *Biochem. Cell Biol.* 79, 253–66.
- Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi:10.1038/nrg2626.
- Minakhina, S., Tan, W., and Steward, R. (2011). JAK/STAT and the GATA factor Pannier control hemocyte maturation and differentiation in *Drosophila*. *Dev. Biol.* 352, 308–16. doi:10.1016/j.ydbio.2011.01.035.
- Mishra, A. K., Bargmann, B. O. R., Tsachaki, M., Fritsch, C., and Sprecher, S. G. (2016). Functional genomics identifies regulators of the phototransduction machinery in the

- Drosophila* larval eye and adult ocelli. *Dev. Biol.* 410, 164–177.
doi:10.1016/j.ydbio.2015.12.026.
- Mishra, M., Oke, a., Lebel, C., McDonald, E. C., Plummer, Z., Cook, T. a., et al. (2010). Pph13 and Orthodenticle define a dual regulatory pathway for photoreceptor cell morphogenesis and function. *Development* 137, 2895–2904. doi:10.1242/dev.051722.
- Monaco, G., van Dam, S., Casal Novo Ribeiro, J. L., Larbi, A., and de Magalhães, J. P. (2015). A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *BMC Evol. Biol.* 15, 259. doi:10.1186/s12862-015-0534-7.
- Montanucci, L., Laayouni, H., Dall’Olio, G. M., and Bertranpetit, J. (2011). Molecular evolution and network-level analysis of the N-glycosylation metabolic pathway across primates. *Mol. Biol. Evol.* 28, 813–23. doi:10.1093/molbev/msq259.
- Mortazavi, A., Williams, B. a, McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi:10.1038/nmeth.1226.
- Mott, R., Yuan, W., Kaisaki, P., Gan, X., Cleak, J., Edwards, A., et al. (2014). The Architecture of Parent-of-Origin Effects in Mice. *Cell* 156, 332–342. doi:10.1016/j.cell.2013.11.043.
- Munger, S. C., Raghupathy, N., Choi, K., Simons, a. K., Gatti, D. M., Hinerfeld, D. a., et al. (2014). RNA-Seq Alignment to Individualized Genomes Improves Transcript Abundance Estimates in Multiparent Populations. *Genetics* 198, 59–73. doi:10.1534/genetics.114.165886.
- Murakami, S., Umetsu, D., Maeyama, Y., Sato, M., Yoshida, S., and Tabata, T. (2007). Focal adhesion kinase controls morphogenesis of the *Drosophila* optic stalk. *Development* 134, 1539–48. doi:10.1242/dev.001529.
- Murali, T., Pacifico, S., Yu, J., Guest, S., Roberts, G. G., and Finley, R. L. (2011). DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res.* 39, D736–43. doi:10.1093/nar/gkq1092.
- Musser, J. M., and Wagner, G. P. (2015). Character trees from transcriptome data: Origin and individuation of morphological characters and the so-called “species signal.” *J. Exp. Zool. Part B Mol. Dev. Evol.* 324, 588–604. doi:10.1002/jez.b.22636.
- Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., et al. (2000). A whole-genome assembly of *Drosophila*. *Science* 287, 2196–204.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–9. doi:10.1126/science.1158441.
- Nguyen, H. T., Bodmer, R., Abmayr, S. M., McDermott, J. C., and Spoerel, N. A. (1994). D-mef2: a *Drosophila* mesoderm-specific MADS box-containing gene with a biphasic expression profile during embryogenesis. *Proc. Natl. Acad. Sci. U. S. A.* 91, 7520–4.
- Nolte, V., Pandey, R. V., Kofler, R., and Schlotterer, C. (2013). Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. *Genome Res.* 23, 99–110. doi:10.1101/gr.139873.112.

- Noordermeer, D., and Duboule, D. (2013). Chromatin looping and organization at developmentally regulated gene loci. *Wiley Interdiscip. Rev. Dev. Biol.* 2, 615–30. doi:10.1002/wdev.103.
- Nunes, M. D. S., Arif, S., Schlötterer, C., and McGregor, A. P. (2013). A perspective on micro-evo-devo: progress and potential. *Genetics* 195, 625–34. doi:10.1534/genetics.113.156463.
- Nüsslein-Volhard, C., and Wieschaus, E. (1980). Mutations affecting segment number and polarity in *Drosophila*. *Nature* 287, 795–801. doi:10.1038/287795a0.
- O'Neill, E. M., Rebay, I., Tjian, R., and Rubin, G. M. (1994). The activities of two Ets-related transcription factors required for *Drosophila* eye development are modulated by the Ras/MAPK pathway. *Cell* 78, 137–147. doi:10.1016/0092-8674(94)90580-0.
- Olsson, L., Levit, G. S., and Hossfeld, U. (2010). Evolutionary developmental biology: its concepts and history with a focus on Russian and German contributions. *Naturwissenschaften* 97, 951–69. doi:10.1007/s00114-010-0720-9.
- Oros, S. M., Tare, M., Kango-Singh, M., and Singh, A. (2010). Dorsal eye selector pannier (pnr) suppresses the eye fate to define dorsal margin of the *Drosophila* eye. *Dev. Biol.* 346, 258–71. doi:10.1016/j.ydbio.2010.07.030.
- Oshlack, A., and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* 4, 14. doi:10.1186/1745-6150-4-14.
- Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98. doi:10.1038/nrg2934.
- Pai, C. Y., Kuo, T. S., Jaw, T. J., Kurant, E., Chen, C. T., Bessarab, D. A., et al. (1998). The Homothorax homeoprotein activates the nuclear localization of another homeoprotein, extradenticle, and suppresses eye development in *Drosophila*. *Genes Dev.* 12, 435–46.
- Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., et al. (2007). Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* 1, 299–312. doi:10.1016/j.stem.2007.08.003.
- Papatsenko, D., Sheng, G., and Desplan, C. (1997). A new rhodopsin in R8 photoreceptors of *Drosophila*: evidence for coordinate expression with Rh3 in R7 cells. *Development* 124, 1665–1673.
- Paris, M., Kaplan, T., Li, X. Y., Villalta, J. E., Lott, S. E., and Eisen, M. B. (2013). Extensive Divergence of Transcription Factor Binding in *Drosophila* Embryos with Highly Conserved Gene Expression. *PLoS Genet.* 9. doi:10.1371/journal.pgen.1003748.
- Perry, G. H., Melsted, P., Marioni, J. C., Wang, Y., Bainer, R., Pickrell, J. K., et al. (2012). Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* 22, 602–10. doi:10.1101/gr.130468.111.
- Pfeiffer, B. D., Jenett, A., Hammonds, A. S., Ngo, T.-T. B., Misra, S., Murphy, C., et al. (2008). Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 105, 9715–20. doi:10.1073/pnas.0803697105.
- Pichaud, F., and Casares, F. (2000). Homothorax and iroquois-C genes are required for the establishment of territories within the developing eye disc. *Mech. Dev.* 96, 15–25. doi:10.1016/S0925-4773(00)00372-5.

- St. Pierre, S. E., Ponting, L., Stefancsik, R., and McQuilton, P. (2014). FlyBase 102 - Advanced approaches to interrogating FlyBase. *Nucleic Acids Res.* 42, 780–788. doi:10.1093/nar/gkt1092.
- Pinnell, J., Lindeman, P. S., Colavito, S., Lowe, C., and Savage, R. M. (2006). The divergent roles of the segmentation gene hunchback. *Integr. Comp. Biol.* 46, 519–32. doi:10.1093/icb/icj054.
- Pinsonneault, R. L., Mayer, N., Mayer, F., Tegegn, N., and Bainton, R. J. (2011). Novel models for studying the blood-brain and blood-eye barriers in *Drosophila*. *Methods Mol. Biol.* 686, 357–69. doi:10.1007/978-1-60761-938-3_17.
- Poelchau, M., Childers, C., Moore, G., Tsavatapalli, V., Evans, J., Lee, C.-Y., et al. (2014). The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Res.* 43, D714–D719. doi:10.1093/nar/gku983.
- Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., et al. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 122, 517–27. doi:10.1016/j.cell.2005.06.026.
- Posnien, N., Hopfen, C., Hilbrant, M., Ramos-Womack, M., Murat, S., Schönauer, A., et al. (2012). Evolution of Eye Morphology and Rhodopsin Expression in the *Drosophila melanogaster* Species Subgroup. *PLoS One* 7, e37346. doi:10.1371/journal.pone.0037346.
- Potier, D., Davie, K., Hulselmans, G., Naval Sanchez, M., Haagen, L., Huynh-Thu, V. A., et al. (2014a). Mapping Gene Regulatory Networks in *Drosophila* Eye Development by Large-Scale Transcriptome Perturbations and Motif Inference. *Cell Rep.* 9, 2290–2303. doi:10.1016/j.celrep.2014.11.038.
- Potier, D., Seyres, D., Guichard, C., Iche-Torres, M., Aerts, S., Herrmann, C., et al. (2014b). Identification of cis-regulatory modules encoding temporal dynamics during development. *BMC Genomics* 15, 534. doi:10.1186/1471-2164-15-534.
- Price, J. T., Tiganis, T., Agarwal, A., Djakiew, D., and Thompson, E. W. (1999). Epidermal growth factor promotes MDA-MB-231 breast cancer cell migration through a phosphatidylinositol 3'-kinase and phospholipase C-dependent mechanism. *Cancer Res.* 59, 5475–8.
- Prud'homme, B., Gompel, N., and Carroll, S. B. (2007). Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104 Suppl , 8605–12. doi:10.1073/pnas.0700488104.
- Punzo, C., Plaza, S., Seimiya, M., Schnupf, P., Kurata, S., Jaeger, J., et al. (2004). Functional divergence between eyeless and twin of eyeless in *Drosophila melanogaster*. *Development* 131, 3943–53. doi:10.1242/dev.01278.
- Quiring, R., Walldorf, U., Kloter, U., and Gehring, W. J. (1994). Homology of the eyeless gene of *Drosophila* to the Small eye gene in mice and Aniridia in humans. *Science* 265, 785–9.
- R Core Team (2015). R: A Language and Environment for Statistical Computing.
- Raff, R., and Kaufman, T. (1983). *Embryos, genes and evolution: the developmental-genetic basis of evolutionary change*. Macmillan, New York.
- Raissig, M. T., Baroux, C., and Grossniklaus, U. (2011). Regulation and flexibility of genomic imprinting during seed development. *Plant Cell* 23, 16–26. doi:10.1105/tpc.110.081018.

- Ramain, P., Heitzler, P., Haenlin, M., and Simpson, P. (1993). *pannier*, a negative regulator of achaete and scute in *Drosophila*, encodes a zinc finger protein with homology to the vertebrate transcription factor GATA-1. *Development* 119, 1277–91.
- Ranade, S. S., Yang-Zhou, D., Kong, S. W., McDonald, E. C., Cook, T. a., and Pignoni, F. (2008). Analysis of the Otd-dependent transcriptome supports the evolutionary conservation of CRX/OTX/OTD functions in flies and vertebrates. *Dev. Biol.* 315, 521–534. doi:10.1016/j.ydbio.2007.12.017.
- Ranganayakulu, G., Elliott, D. A., Harvey, R. P., and Olson, E. N. (1998). Divergent roles for NK-2 class homeobox genes in cardiogenesis in flies and mice. *Development* 125, 3037–48.
- Rangarajan, R., Courvoisier, H., and Gaul, U. (2001). Dpp and Hedgehog mediate neuron-glia interactions in *Drosophila* eye development by promoting the proliferation and motility of subretinal glia. *Mech. Dev.* 108, 93–103. doi:S0925477301005019.
- Rangarajan, R., Gong, Q., and Gaul, U. (1999). Migration and function of glia in the developing *Drosophila* eye. *Development* 126, 3285–3292.
- Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., et al. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* 14, R95. doi:10.1186/gb-2013-14-9-r95.
- Rau, A., Gallopin, M., Celeux, G., and Jaffrézic, F. (2013). Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* 29, 2146–52. doi:10.1093/bioinformatics/btt350.
- Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., and Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics* 31, 1420–1427. doi:10.1093/bioinformatics/btu845.
- Ready, D. F., Hanson, T. E., and Benzer, S. (1976). Development of the *Drosophila* retina, a neurocrystalline lattice. *Dev. Biol.* 53, 217–40.
- Reilly, J. G., and Thomas, C. A. (1980). Length polymorphisms, restriction site variation, and maternal inheritance of mitochondrial DNA of *Drosophila melanogaster*. *Plasmid* 3, 109–115. doi:10.1016/0147-619X(80)90102-X.
- Reinke, R., and Zipursky, S. L. (1988). Cell-cell interaction in the *Drosophila* retina: the bride of sevenless gene is required in photoreceptor cell R8 for R7 cell development. *Cell* 55, 321–30.
- Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., et al. (2005). Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res.* 15, 1–18. doi:10.1101/gr.3059305.
- Rieckhof, G. E., Casares, F., Ryoo, H. D., Abu-Shaar, M., and Mann, R. S. (1997). Nuclear translocation of extradenticle requires homothorax, which encodes an extradenticle-related homeodomain protein. *Cell* 91, 171–83.
- Rifkin, S. A., Kim, J., and White, K. P. (2003). Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* 33, 138–44. doi:10.1038/ng1086.
- Rister, J., and Desplan, C. (2011). The retinal mosaics of opsin expression in invertebrates and vertebrates. *Dev. Neurobiol.* 71, 1212–26. doi:10.1002/dneu.20905.

- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007.
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi:10.1038/nbt.1754.
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–40. doi:10.1093/bioinformatics/btp616.
- Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25. doi:10.1186/gb-2010-11-3-r25.
- Rokas, A. (2008). The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu. Rev. Genet.* 42, 235–51. doi:10.1146/annurev.genet.42.110807.091513.
- Ronshaugen, M., McGinnis, N., and McGinnis, W. (2002). Hox protein mutation and macroevolution of the insect body plan. *Nature* 415, 914–7. doi:10.1038/nature716.
- Roux, J., Rosikiewicz, M., and Robinson-Rechavi, M. (2015). What to compare and how: Comparative transcriptomics for Evo-Devo. *J. Exp. Zool. B. Mol. Dev. Evol.* 324, 372–82. doi:10.1002/jez.b.22618.
- Royet, J., and Finkelstein, R. (1995). Pattern formation in *Drosophila* head development: the role of the orthodenticle homeobox gene. *Development* 121, 3561–3572.
- Royet, J., and Finkelstein, R. (1996). hedgehog, wingless and orthodenticle specify adult head development in *Drosophila*. *Development* 122, 1849–1858.
- Russo, F., and Angelini, C. (2014). RNASEqGUI: a GUI for analysing RNA-Seq data. *Bioinformatics*, 1–3. doi:10.1093/bioinformatics/btu308.
- Sanes, J. R., and Zipursky, S. L. (2010). Design principles of insect and vertebrate visual systems. *Neuron* 66, 15–36. doi:10.1016/j.neuron.2010.01.018.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–7.
- Dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., et al. (2014). FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.* doi:10.1093/nar/gku1099.
- Sato, A., and Tomlinson, A. (2007). Dorsal-ventral midline signaling in the developing *Drosophila* eye. *Development* 134, 659–667. doi:10.1242/dev.02786.
- Satya, R. V., Zavaljevski, N., and Reifman, J. (2012). A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Res.* 40, e127–e127. doi:10.1093/nar/gks425.
- Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods* 9, 676–82. doi:10.1038/nmeth.2019.
- Schmidt-Ott, U., González-Gaitán, M., Jäckle, H., and Technau, G. M. (1994). Number, identity, and sequence of the *Drosophila* head segments as revealed by neural

- elements and their deletion patterns in mutants. *Proc. Natl. Acad. Sci. U. S. A.* 91, 8363–7.
- Schmidt-Ott, U., González-Gaitán, M., and Technau, G. M. (1995). Analysis of neural elements in head-mutant *Drosophila* embryos suggests segmental origin of the optic lobes. *Roux's Arch. Dev. Biol.* 205, 31–44. doi:10.1007/BF00188841.
- Schmucker, D., Jäckle, H., and Gaul, U. (1997). Genetic analysis of the larval optic nerve projection in *Drosophila*. *Development* 124, 937–48.
- Schröder, R. (2003). The genes orthodenticle and hunchback substitute for bicoid in the beetle *Tribolium*. *Nature* 422, 621–5. doi:10.1038/nature01536.
- Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28, 1086–1092. doi:10.1093/bioinformatics/bts094.
- Schwabe, T., Bainton, R. J., Fetter, R. D., Heberlein, U., and Gaul, U. (2005). GPCR Signaling Is Required for Blood-Brain Barrier Formation in *Drosophila*. *Cell* 123, 133–144. doi:10.1016/j.cell.2005.08.037.
- Scott, M. P., Tamkun, J. W., and Hartzell, G. W. (1989). The structure and function of the homeodomain. *Biochim. Biophys. Acta* 989, 25–48.
- Seyednasrollah, F., Laiho, A., and Elo, L. L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.*, bbt086–. doi:10.1093/bib/bbt086.
- Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jónsson, B., et al. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 428, 717–23. doi:10.1038/nature02415.
- Sharma, H. S., and Dey, P. K. (1986). Influence of long-term immobilization stress on regional blood-brain barrier permeability, cerebral blood flow and 5-HT level in conscious normotensive young rats. *J. Neurol. Sci.* 72, 61–76.
- Sharma, S. V., Bell, D. W., Settleman, J., and Haber, D. A. (2007). Epidermal growth factor receptor mutations in lung cancer. *Nat. Rev. Cancer* 7, 169–81. doi:10.1038/nrc2088.
- Shcherbata, H. R., Althausen, C., Findley, S. D., and Ruohola-Baker, H. (2004). The mitotic-to-endocycle switch in *Drosophila* follicle cells is executed by Notch-dependent regulation of G1/S, G2/M and M/G1 cell-cycle transitions. *Development* 131, 3169–81. doi:10.1242/dev.01172.
- Shen, Y., Garcia, T., Pabuwal, V., Boswell, M., Pasquali, A., Beldorth, I., et al. (2013). Alternative strategies for development of a reference transcriptome for quantification of allele specific expression in organisms having sparse genomic resources. *Comp. Biochem. Physiol. Part D Genomics Proteomics* 8, 11–16. doi:10.1016/j.cbd.2012.10.006.
- Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145. doi:10.1038/nbt1486.
- Shilo, B. (2003). Signaling by the *Drosophila* epidermal growth factor receptor pathway during development. *Exp. Cell Res.* 284, 140–149. doi:10.1016/S0014-4827(02)00094-0.
- Shilo, B.-Z. (2005). Regulating the dynamics of EGF receptor signaling in space and time. *Development* 132, 4017–27. doi:10.1242/dev.02006.

- Shir-Shapira, H., Sharabany, J., Filderman, M., Ideses, D., Ovadia-Shochat, A., Mannervik, M., et al. (2015). Structure-Function Analysis of the *Drosophila melanogaster* Caudal Transcription Factor Provides Insights into Core Promoter-preferential Activation. *J. Biol. Chem.* 290, 17293–305. doi:10.1074/jbc.M114.632109.
- Siegal, M. L., Promislow, D. E. L., and Bergman, A. (2007). Functional and evolutionary inference in gene networks: does topology matter? *Genetica* 129, 83–103. doi:10.1007/s10709-006-0035-0.
- Silies, M., Yuva, Y., Engelen, D., Aho, a., Stork, T., and Klambt, C. (2007). Glial Cell Migration in the Eye Disc. *J. Neurosci.* 27, 13130–13139. doi:10.1523/JNEUROSCI.3583-07.2007.
- Singh, A., Chan, J., Chern, J. J., and Choi, K.-W. (2005). Genetic interaction of *Lobe* with its modifiers in dorsoventral patterning and growth of the *Drosophila* eye. *Genetics* 171, 169–83. doi:10.1534/genetics.105.044180.
- Singh, A., and Choi, K.-W. (2003). Initial state of the *Drosophila* eye before dorsoventral specification is equivalent to ventral. *Development* 130, 6351–6360. doi:10.1242/dev.00864.
- Sivachenko, A., Li, Y., Abruzzi, K. C., and Rosbash, M. (2013). The transcription factor *Mef2* links the *Drosophila* core clock to *Fas2*, neuronal morphology, and circadian behavior. *Neuron* 79, 281–92. doi:10.1016/j.neuron.2013.05.015.
- Skaar, D. A., and Jirtle, R. L. (2015). Analysis of Imprinted Gene Regulation. *Methods Mol. Biol.* doi:10.1007/7651_2015_264.
- Skultétyová, I., Tokarev, D., and Jezová, D. (1998). Stress-induced increase in blood-brain barrier permeability in control and monosodium glutamate-treated rats. *Brain Res. Bull.* 45, 175–8.
- Slater, G. S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 31. doi:10.1186/1471-2105-6-31.
- Smolla, M., Ruchty, M., Nagel, M., and Kleineidam, C. J. (2014). Clearing pigmented insect cuticle to investigate small insects' organs in situ using confocal laser-scanning microscopy (CLSM). *Arthropod Struct. Dev.* 43, 175–181. doi:10.1016/j.asd.2013.12.006.
- Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* 3, 1–25. doi:10.2202/1544-6115.1027.
- Snodgrass, R. (1935). *Principles of Insect Morphology*. New York: McGraw-Hill Book Co.
- Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14, 91. doi:10.1186/1471-2105-14-91.
- Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 1–15. doi:10.12688/f1000research.7563.1.
- Spitz, F., and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13, 613–626. doi:10.1038/nrg3207.
- Sprecher, S. G., and Desplan, C. (2008). Switch of rhodopsin expression in terminally differentiated *Drosophila* sensory neurons. *Nature* 454, 533–7. doi:10.1038/nature07062.

- Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24, 637–644. doi:10.1093/bioinformatics/btn013.
- Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19, ii215–ii225. doi:10.1093/bioinformatics/btg1080.
- Stanojevic, D., Small, S., and Levine, M. (1991). Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. *Science* 254, 1385–7.
- Steiner, C. C., Weber, J. N., and Hoekstra, H. E. (2007). Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS Biol.* 5, e219. doi:10.1371/journal.pbio.0050219.
- Stern, D. L. (1998). A role of Ultrabithorax in morphological differences between *Drosophila* species. *Nature* 396, 463–466. doi:10.1038/24863.
- Stern, D. L. and Orgogozo, V. (2009). Is genetic evolution predictable? *Science* 323, 746–751.
- Stevenson, K. R., Coolon, J. D., and Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. *BMC Genomics* 14, 536. doi:10.1186/1471-2164-14-536.
- Sturtevant, A. H. (1919). A New Species Closely Resembling *Drosophila Melanogaster*. *Psyche (Stuttg.)* 26, 153–155.
- Sturtevant, A. H. (1939). On the Subdivision of the Genus *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 25, 137–41.
- Sucena, E., and Stern, D. L. (2000). Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc. Natl. Acad. Sci. U. S. A.* 97, 4530–4.
- Sugano, S. S., Shirakawa, M., Takagi, J., Matsuda, Y., Shimada, T., Hara-Nishimura, I., et al. (2014). CRISPR/Cas9-mediated targeted mutagenesis in the liverwort *Marchantia polymorpha* L. *Plant Cell Physiol.* 55, 475–81. doi:10.1093/pcp/pcu014.
- Suvorov, A., Nolte, V., Pandey, R. V., Franssen, S. U., Futschik, A., and Schlötterer, C. (2013). Intra-Specific Regulatory Variation in *Drosophila pseudoobscura*. *PLoS One* 8, e83547. doi:10.1371/journal.pone.0083547.
- Tahayato, A., Sonnevile, R., Pichaud, F., Wernet, M. F., Papatsenko, D., Beaufils, P., et al. (2003). Otd/Crx, a dual regulator for the specification of ommatidia subtypes in the *Drosophila* retina. *Dev. Cell* 5, 391–402. doi:10.1016/S1534-5807(03)00239-9.
- Talarico, F., Brandmayr, P., Giglio, A., Massolo, A., and Brandmayr, T. Z. (2011). Morphometry of eyes, antennae and wings in three species of *Siagona* (Coleoptera, Carabidae). *Zookeys*, 203–14. doi:10.3897/zookeys.100.1528.
- Tallafuss, A., and Bally-Cuif, L. (2002). Formation of the head-trunk boundary in the animal body plan: an evolutionary perspective. *Gene* 287, 23–32.
- Tanaka, K. M., Hopfen, C., Herbert, M. R., Schlötterer, C., Stern, D. L., Masly, J. P., et al. (2015). Genetic architecture and functional characterization of genes underlying the rapid diversification of male external genitalia between *Drosophila simulans* and *Drosophila mauritiana*. *Genetics* 200, 357–69. doi:10.1534/genetics.114.174045.

- Tautz, D., Lehmann, R., Schnurch, H., Schuh, R., Seifert, E., Kienlin, A., et al. (1987). Finger protein of novel structure encoded by hunchback, a second member of the gap class of *Drosophila* segmentation genes. *Nature* 327, 383–389.
- Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi:10.1093/bib/bbs017.
- Tirosch, I., Reikhav, S., Levy, A. A., and Barkai, N. (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* 324, 659–62. doi:10.1126/science.1169766.
- Tomlinson, a, and Ready, D. F. (1987). Cell fate in the *Drosophila* ommatidium. *Dev. Biol.* 123, 264–275. doi:10.1016/0012-1606(87)90448-9.
- Torres-Oliva, M., Almudi, I., McGregor, A. P., and Posnien, N. A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species. (in revision in BMC Genomics).
- Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–11. doi:10.1093/bioinformatics/btp120.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–78. doi:10.1038/nprot.2012.016.
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M. J. van, et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi:10.1038/nbt.1621.
- Treisman, J. E. (2013). Retinal differentiation in *Drosophila*. *Wiley Interdiscip. Rev. Dev. Biol.* 2, 545–557. doi:10.1002/wdev.100.
- Treisman, J. E., and Rubin, G. M. (1995). wingless inhibits morphogenetic furrow movement in the *Drosophila* eye disc. *Development* 121, 3519–27.
- True, J. R., and Haag, E. S. (2001). Developmental system drift and flexibility in evolutionary trajectories. *Evol. Dev.* 3, 109–19.
- Tsacas, L., and David, J. R. (1974). *Drosophila mauritiana* n.sp. du groupe melanogaster de l'île Maurice. *Bull. Soc. ent. Fr.* 79, 42–46.
- Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., et al. (2009). FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res.* 37, D555–559. doi:10.1093/nar/gkn788.
- Ulitsky, I., and Shamir, R. (2007). Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.* 1, 8. doi:10.1186/1752-0509-1-8.
- Unhavaithaya, Y., and Orr-weaver, T. L. (2012). Polyploidization of glia in neural development links tissue growth to blood – brain barrier integrity. *Genes Dev.*, 31–36. doi:10.1101/gad.177436.111.Freely.
- Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., and Leunissen, J. A. M. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* 35, W71–4. doi:10.1093/nar/gkm306.

- Vandendries, E. R., Johnson, D., and Reinke, R. (1996). orthodenticle is required for photoreceptor cell development in the *Drosophila* eye. *Dev Biol* 173, 243–255. doi:10.1006/dbio.1996.0020 [pii].
- Waddington, C. H. (1961). *New Patterns in Genetics and Development*. Columbia Univ. Press, New York/London.
- Waddington, C. H., and Perry, M. M. (1960). The Ultra-Structure of the Developing Eye of *Drosophila*. *Proc. R. Soc. B Biol. Sci.* 153, 155–178. doi:10.1098/rspb.1960.0094.
- Wagner, G. P. (2007). The developmental genetics of homology. *Nat. Rev. Genet.* 8, 473–9. doi:10.1038/nrg2099.
- Wagner, G. P., Kin, K., and Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131, 281–285. doi:10.1007/s12064-012-0162-3.
- Wagner, G. P., and Lynch, V. J. (2008). The gene regulatory logic of transcription factor evolution. *Trends Ecol. Evol.* 23, 377–85. doi:10.1016/j.tree.2008.03.006.
- Walker, A., and Parkhill, J. (2008). Single-cell genomics. *Nat. Rev. Microbiol.* 6, 176–7. doi:10.1038/nrmicro1862.
- Wang, Y., Li, Z., Xu, J., Zeng, B., Ling, L., You, L., et al. (2013). The CRISPR/Cas system mediates efficient genome engineering in *Bombyx mori*. *Cell Res.* 23, 1414–6. doi:10.1038/cr.2013.146.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484.
- Warnefors, M., and Kaessmann, H. (2013). Evolution of the correlation between expression divergence and protein divergence in mammals. *Genome Biol. Evol.* 5, 1324–35. doi:10.1093/gbe/evt093.
- Warrant, E. J., and McIntyre, P. D. (1993). Arthropod eye design and the physical limits to spatial resolving power. *Prog. Neurobiol.* 40, 413–461. doi:10.1016/0301-0082(93)90017-M.
- Weasner, B. M., and Kumar, J. P. (2013). Competition among gene regulatory networks imposes order within the eye-antennal disc of *Drosophila*. *Development* 140, 205–215. doi:10.1242/dev.085423.
- Weber, U., Pataki, C., Mihaly, J., and Mlodzik, M. (2008). Combinatorial signaling by the Frizzled/PCP and Egfr pathways during planar cell polarity establishment in the *Drosophila* eye. *Dev. Biol.* 316, 110–23. doi:10.1016/j.ydbio.2008.01.016.
- Wedd, L., Kucharski, R., and Maleszka, R. (2015). Differentially methylated obligatory epialleles modulate context-dependent LAM gene expression in the honeybee *Apis mellifera*. *Epigenetics*, 1–10. doi:10.1080/15592294.2015.1107695.
- Wei, X., and Wang, X. (2013). A computational workflow to identify allele-specific expression and epigenetic modification in maize. *Genomics. Proteomics Bioinformatics* 11, 247–52. doi:10.1016/j.gpb.2013.05.006.
- Wernet, M. F., Mazzoni, E. O., Çelik, A., Duncan, D. M., Duncan, I., and Desplan, C. (2006). Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature* 440, 174–180. doi:10.1038/nature04615.
- Wetterstrand, K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcosts. Accessed 03/2016.

- Wiersdorff, V., Lecuit, T., Cohen, S. M., and Mlodzik, M. (1996). Mad acts downstream of Dpp receptors, revealing a differential requirement for dpp signaling in initiation and propagation of morphogenesis in the *Drosophila* eye. *Development* 122, 2153–62.
- Wilczynski, B., and Furlong, E. E. M. M. (2010). Challenges for modeling global gene regulatory networks during development: Insights from *Drosophila*. *Dev. Biol.* 340, 161–169. doi:10.1016/j.ydbio.2009.10.032.
- Williams, C. R., Baccarella, A., Parrish, J. Z., and Kim, C. C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* 17, 103. doi:10.1186/s12859-016-0956-2.
- Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L. J., et al. (2008). Species-specific transcription in mice carrying human chromosome 21. *Science* 322, 434–8. doi:10.1126/science.1160930.
- Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88. doi:10.1038/nature02698.
- Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2008). Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat. Genet.* 40, 346–350. doi:10.1038/ng.77.
- Wolf, J. B. W., Lindell, J., and Backström, N. (2010). Speciation genetics: current status and evolving approaches. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 365, 1717–33. doi:10.1098/rstb.2010.0023.
- Wolff, T., and Ready, D. F. (1993). *Pattern formation in the Drosophila retina*. In: *The Development of Drosophila melanogaster*. Cold Spring Harbor Laboratory Press.
- Wray, G. a (2007). The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8, 206–16. doi:10.1038/nrg2063.
- Yamaguchi, S., Desplan, C., and Heisenberg, M. (2010). Contribution of photoreceptor subtypes to spectral wavelength preference in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 107, 5634–9. doi:10.1073/pnas.0809398107.
- Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., and Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science* 297, 1143. doi:10.1126/science.1072545.
- Yandell, M., and Ence, D. (2012). A beginner’s guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–342. doi:10.1038/nrg3174.
- Yáñez-Cuna, J. O., Kvon, E. Z., and Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. *Trends Genet.* 29, 11–22. doi:10.1016/j.tig.2012.09.007.
- Yewale, C., Baradia, D., Vhora, I., Patil, S., and Misra, A. (2013). Epidermal growth factor receptor targeting in cancer: a review of trends and strategies. *Biomaterials* 34, 8690–707. doi:10.1016/j.biomaterials.2013.07.100.
- Yuva-Aydemir, Y., Bauke, A.-C., and Klämbt, C. (2011). Spinster controls Dpp signaling during glial migration in the *Drosophila* eye. *J. Neurosci.* 31, 7005–7015. doi:10.1523/JNEUROSCI.0459-11.2011.
- Zhang, X., and Borevitz, J. O. (2009). Global analysis of allele-specific expression in *Arabidopsis thaliana*. *Genetics* 182, 943–54. doi:10.1534/genetics.109.103499.
- Zhao, L., Wit, J., Svetec, N., and Begun, D. J. (2015). Parallel Gene Expression Differences between Low and High Latitude Populations of *Drosophila melanogaster* and *D. simulans*. *PLOS Genet.* 11, e1005184. doi:10.1371/journal.pgen.1005184.

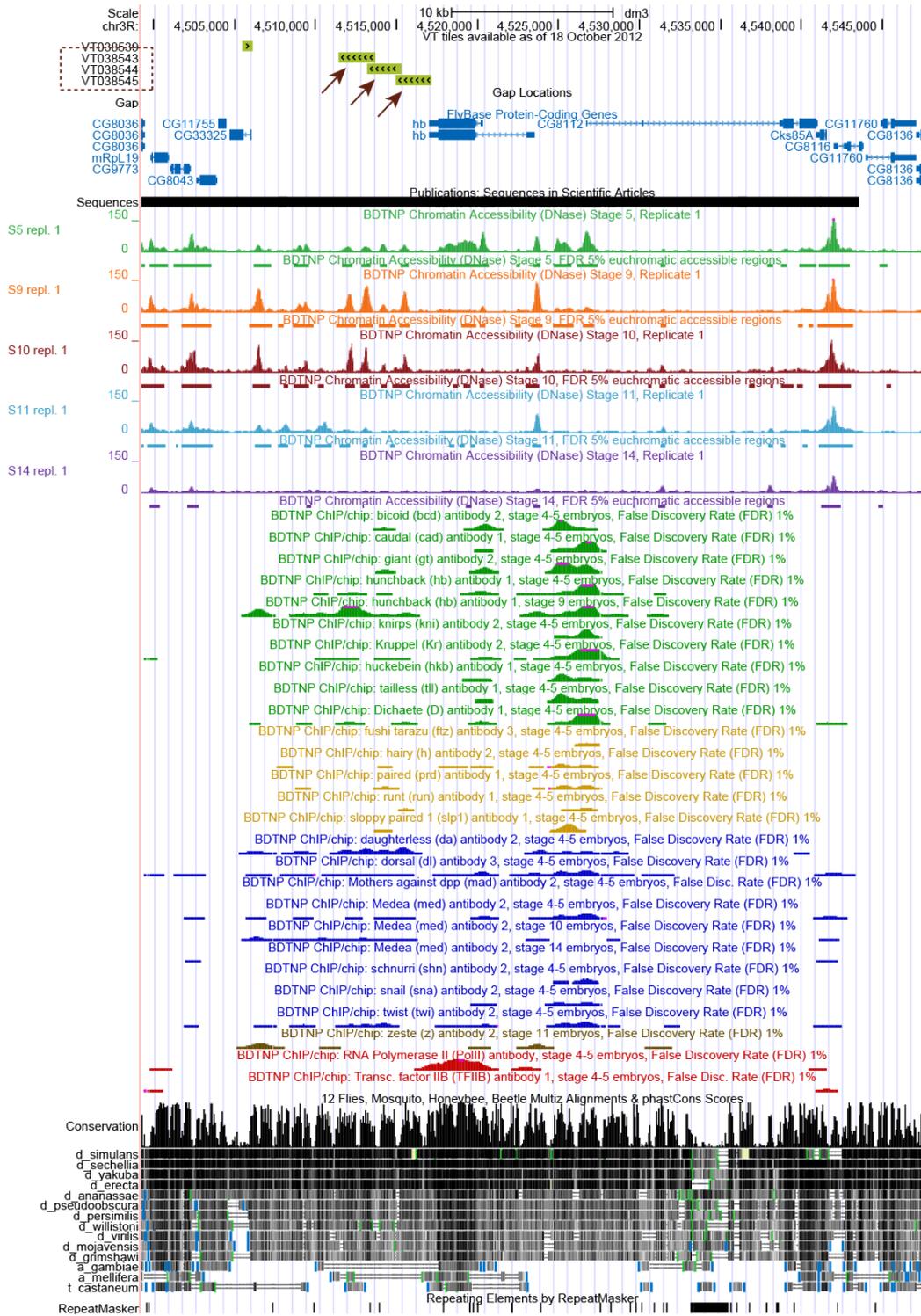
- Zhao, S., and Zhang, B. (2015). A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics* 16, 97. doi:10.1186/s12864-015-1308-8.
- Zheng, L., Zhang, J., and Carthew, R. W. (1995). *frizzled* regulates mirror-symmetric pattern formation in the *Drosophila* eye. *Development* 121, 3045–55.
- Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* 462, 65–70. doi:10.1038/nature08531.
- Zuker, C. S., Cowman, A. F., and Rubin, G. M. (1985). Isolation and structure of a rhodopsin gene from *D. melanogaster*. *Cell* 40, 851–8.

7 Appendix

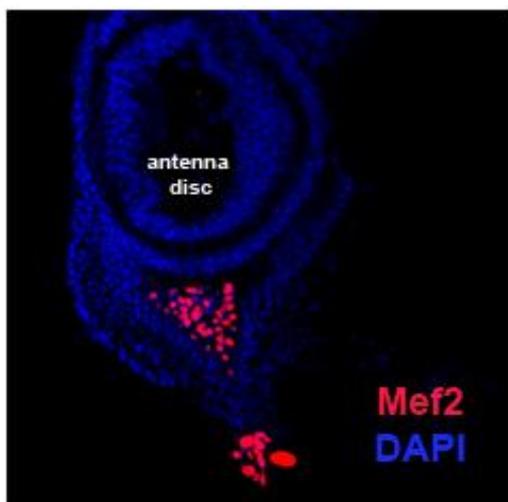
7.1 Abbreviations

AEL	After Egg Laying
LI	1 st larval instar
LII	2 nd larval instar
LIII	3 rd larval instar
<i>D. mel</i>	<i>Drosophila melanogaster</i>
<i>D. sim</i>	<i>Drosophila simulans</i>
<i>D. mau</i>	<i>Drosophila mauritiana</i>
YVF	<i>yellow vermillion forked</i>
OreR	OregonR
GO	Gene Ontology
dsRNA	double-stranded RNA
ASE	Allele-Specific Expression
TF	Transcription Factor
UTR	Untranslated region
CRE	<i>cis</i> -regulatory element
GRN	Gene regulatory network
QTL	Quantitative trait loci
RPKM	Reads per kilobase per million
SE	Single-End
PE	Paired-End

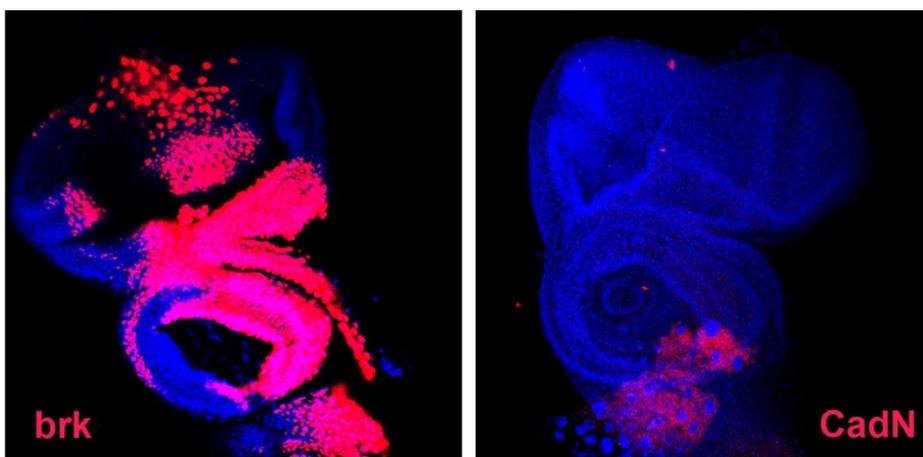
7.2 Supplementary Figures



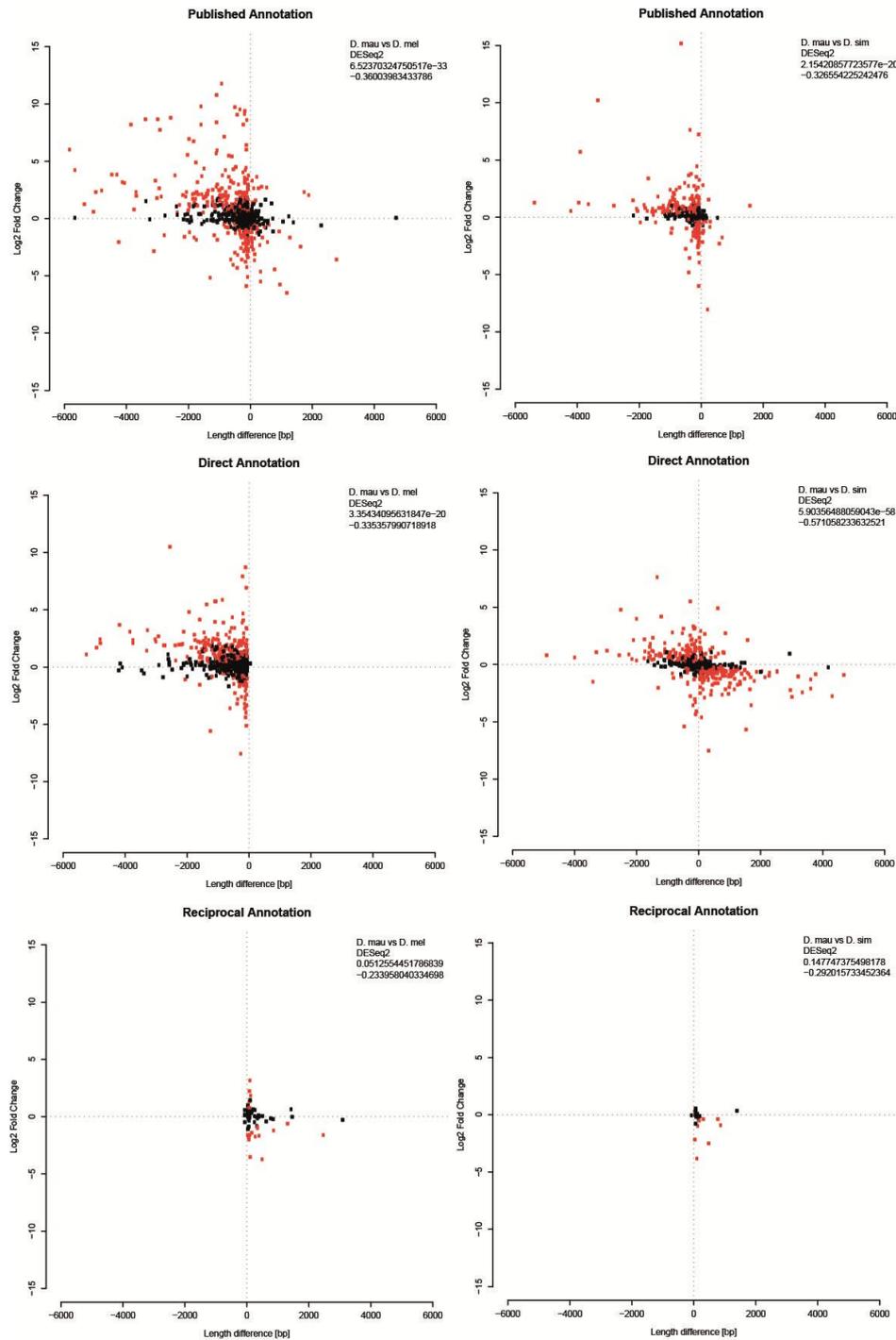
Supplementary Figure 1. Genomic location of Vienna Tile Hb driver lines. Arrows indicate the regions used to drive *hb* expression with Gal4 system. Below, are colored tracks provided by the BDTNP project (Li et al., 2008) showing open chromatin profiles and transcription factor binding. The last black tracks show sequence conservation across different insect species. These tracks were visualized using UCSC Browser (Kent et al., 2002).



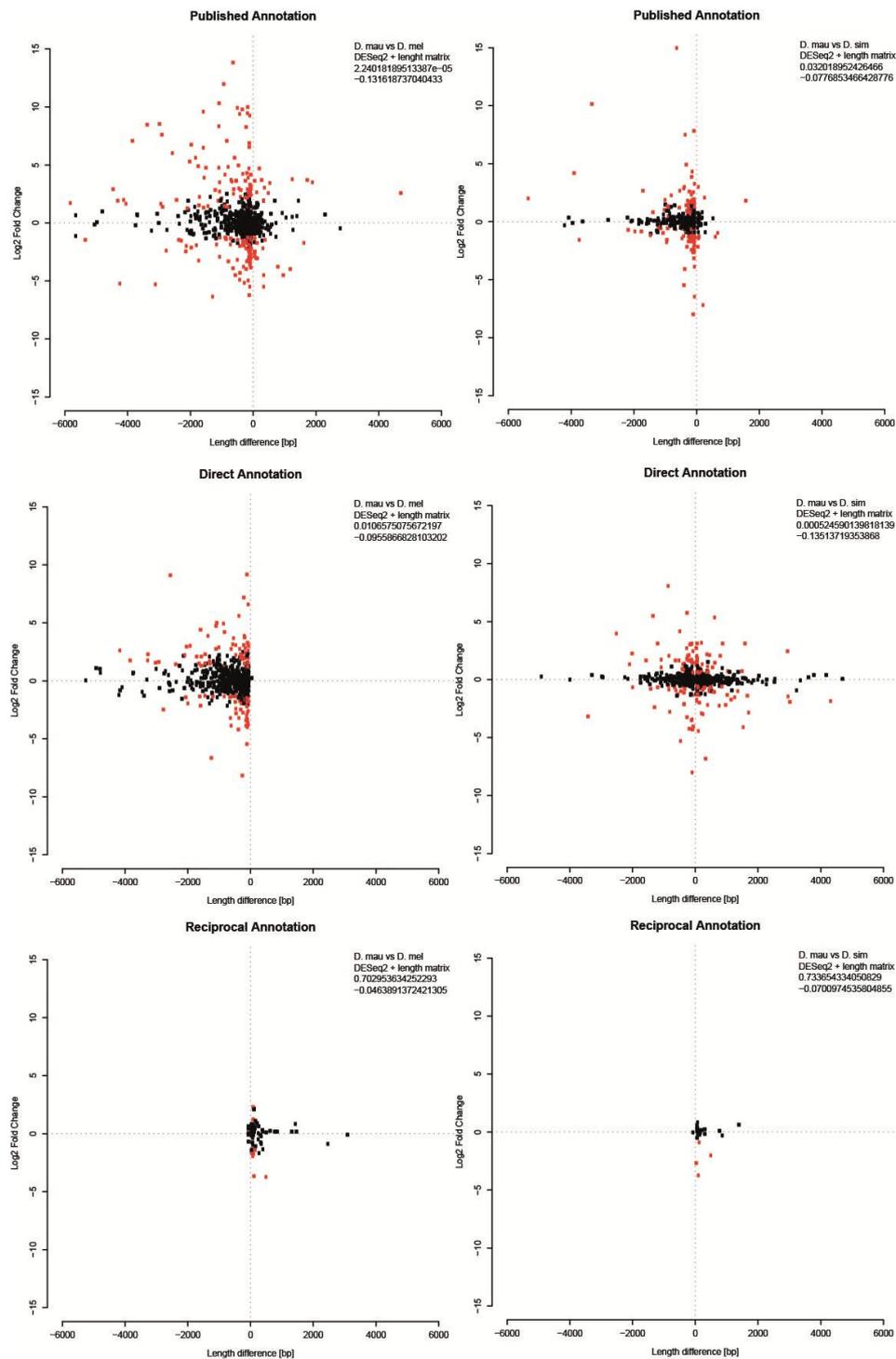
Supplementary Figure 2. *mef2* expression



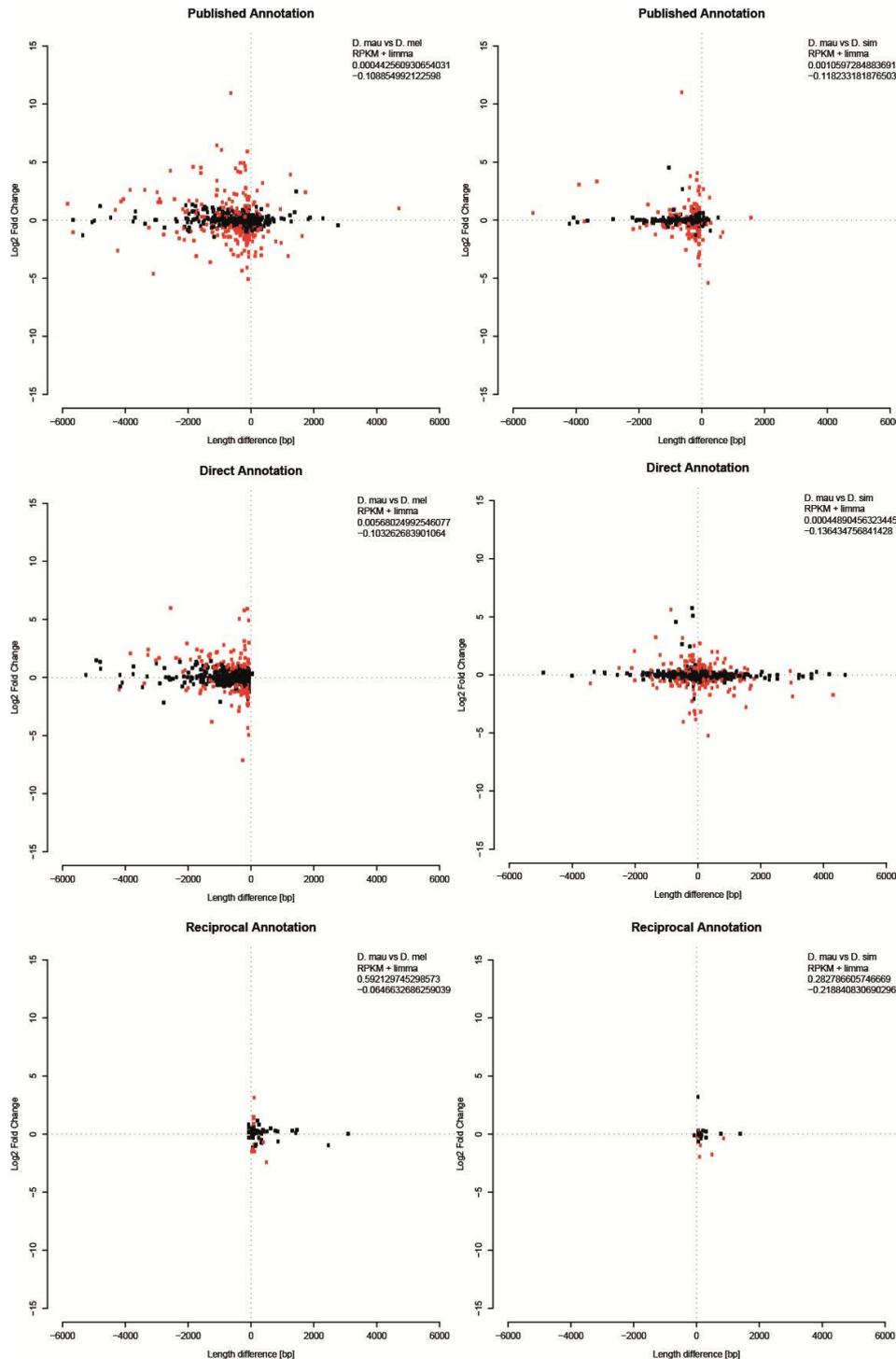
Supplementary Figure 3. *brk* and *CadN* expression



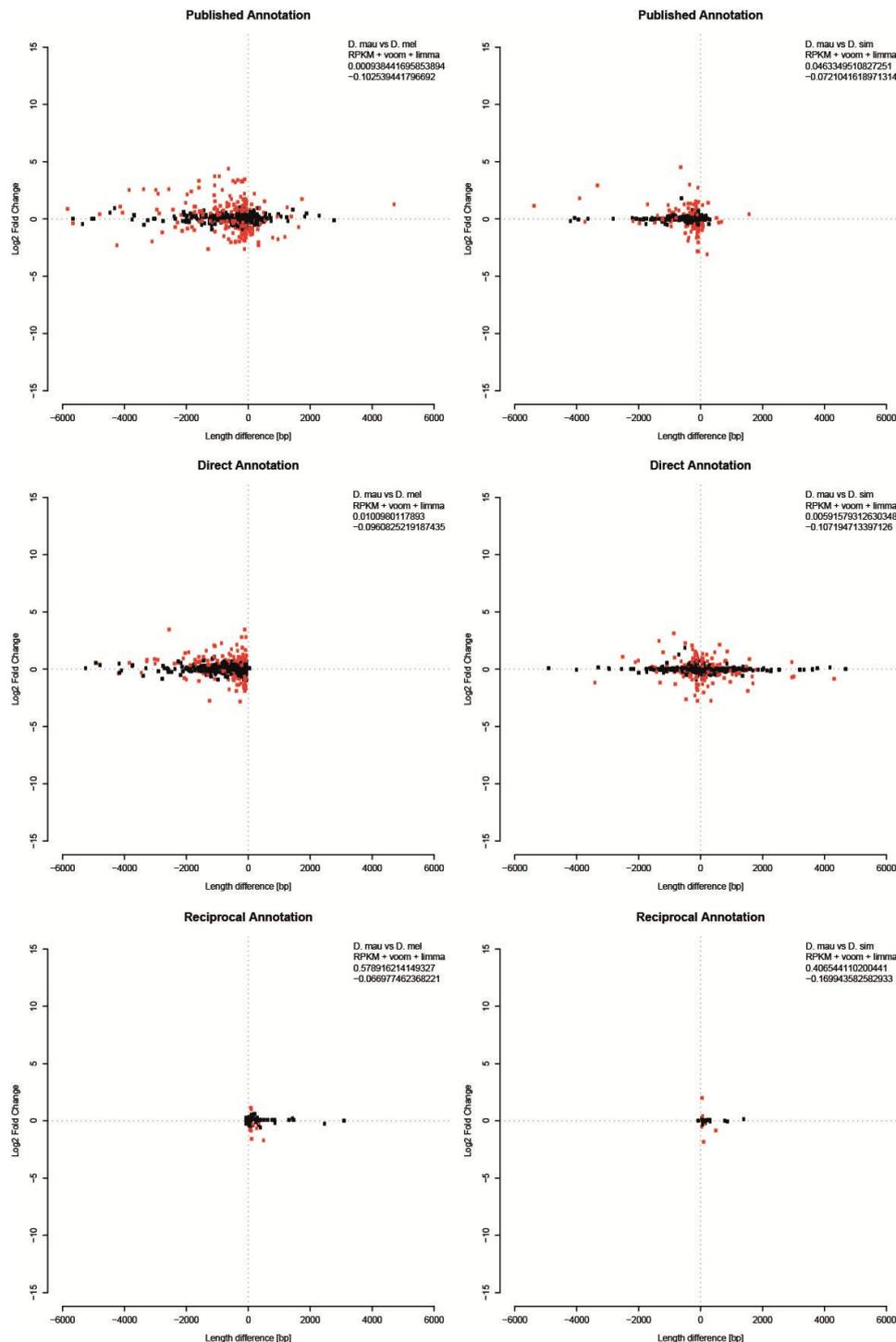
Supplementary Figure 4. Correlation plots for DESeq2 using direct counts. Relation between length differences and the log₂-fold change. Comparisons between *D. mauritiana* and *D. melanogaster* are shown on the left side, comparisons between *D. mauritiana* and *D. simulans* are shown on the right side. On the first row, the published annotations are used as mapping references; on the second row, the directly re-annotated references are used as mapping references and on the third row, the reciprocally re-annotated references are used. Dots represent genes with length difference > 49 bp in these annotations. Genes significantly differentially expressed in the presented analysis ($p\text{-adj} < 0.05$) are shown in red. A negative log₂-fold change indicates higher expression in *D. mauritiana*. A positive length difference indicates that the ortholog of *D. mauritiana* is longer. The p -value and rho of the Spearman's rank correlation are indicated on the upper right side of the plots.



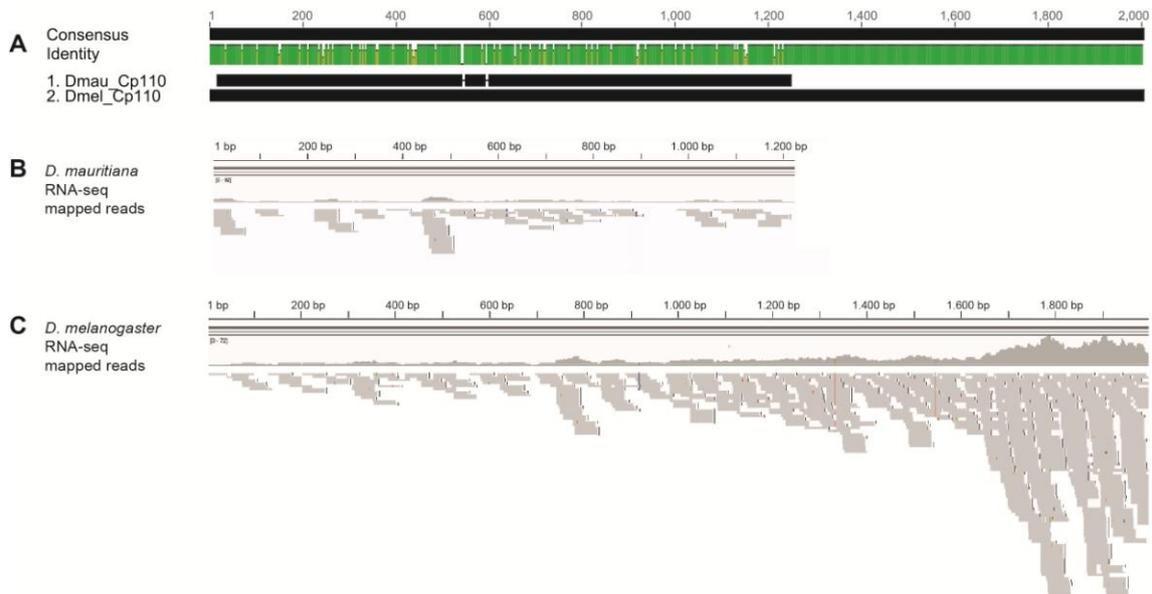
Supplementary Figure 5. Correlation plots for DESeq2 including length correction. Relation between length differences and the log₂-fold change. Comparisons between *D. mauritiana* and *D. melanogaster* are shown on the left side, comparisons between *D. mauritiana* and *D. simulans* are shown on the right side. On the first row, the published annotations are used as mapping references; on the second row, the directly re-annotated references are used as mapping references and on the third row, the reciprocally re-annotated references are used. Dots represent genes with length difference > 49 bp in these annotations. Genes significantly differentially expressed in the presented analysis (p-adj < 0.05) are shown in red. A negative log₂-fold change indicates higher expression in *D. mauritiana*. A positive length difference indicates that the ortholog of *D. mauritiana* is longer. The p-value and rho of the Spearman's rank correlation are indicated on the upper right side of the plots.



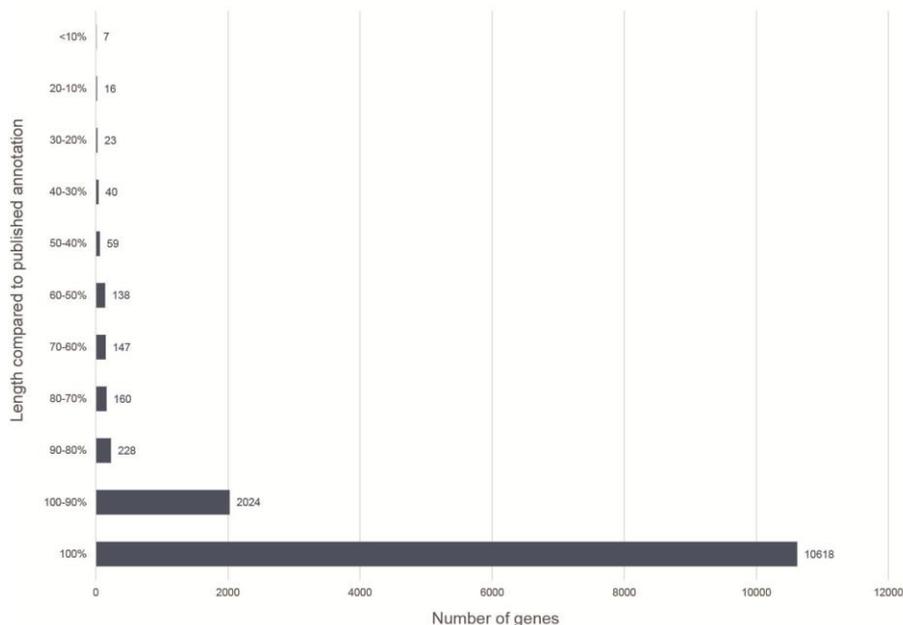
Supplementary Figure 6. Correlation plots for RPKM-limma. Relation between length differences and the log₂-fold change. Comparisons between *D. mauritiana* and *D. melanogaster* are shown on the left side, comparisons between *D. mauritiana* and *D. simulans* are shown on the right side. On the first row, the published annotations are used as mapping references; on the second row, the directly re-annotated references are used as mapping references and on the third row, the reciprocally re-annotated references are used. Dots represent genes with length difference > 49 bp in these annotations. Genes significantly differentially expressed in the presented analysis ($p\text{-adj} < 0.05$) are shown in red. A negative log₂-fold change indicates higher expression in *D. mauritiana*. A positive length difference indicates that the ortholog of *D. mauritiana* is longer. The p-value and rho of the Spearman's rank correlation are indicated on the upper right side of the plots.



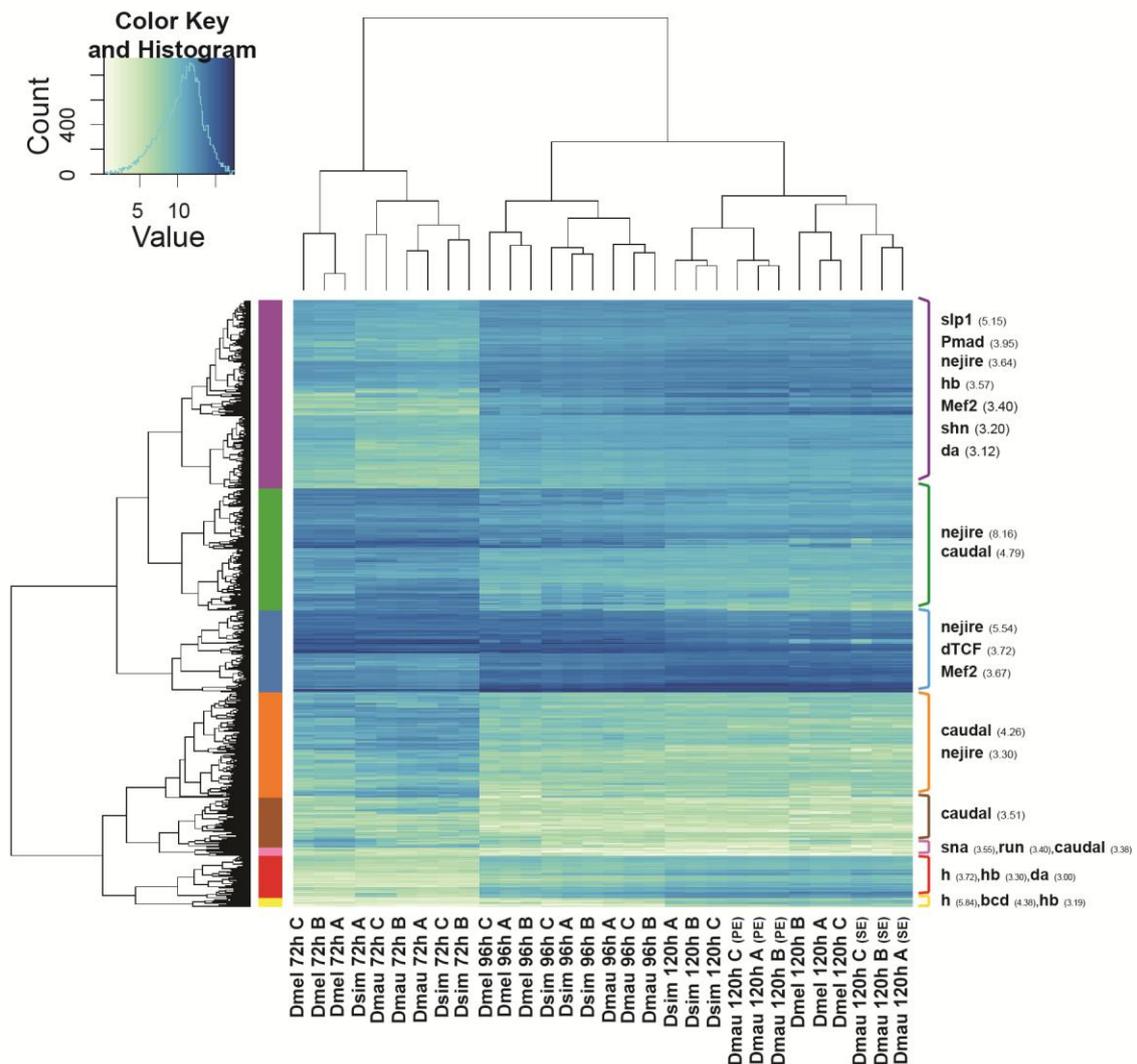
Supplementary Figure 7. Correlation plots for RPKM-voom-limma. Relation between length differences and the log₂-fold change. Comparisons between *D. mauritiana* and *D. melanogaster* are shown on the left side, comparisons between *D. mauritiana* and *D. simulans* are shown on the right side. On the first row, the published annotations are used as mapping references; on the second row, the directly re-annotated references are used as mapping references and on the third row, the reciprocally re-annotated references are used. Dots represent genes with length difference > 49 bp in these annotations. Genes significantly differentially expressed in the presented analysis ($p\text{-adj} < 0.05$) are shown in red. A negative log₂-fold change indicates higher expression in *D. mauritiana*. A positive length difference indicates that the ortholog of *D. mauritiana* is longer. The p -value and rho of the Spearman's rank correlation are indicated on the upper right side of the plots.



Supplementary Figure 8. *Cp110* coverage. (A) Alignment of the published annotated transcripts of the gene *Cp110* in *D. mauritiana* (upper, shorter black bar) and *D. melanogaster* (lower, longer black bar). The top ruler indicates the length of the alignment in bp, the green bar shows the base similarity. (B) *D. mauritiana* RNA-seq reads mapped to the body of the *D. mauritiana* *Cp110* transcript. (C) *D. melanogaster* RNA-seq reads mapped to the body of the *D. melanogaster* *Cp110* transcript. Very few reads map to the 5' region, more reads map from the central portion of the gene, and many more to the 3' end.

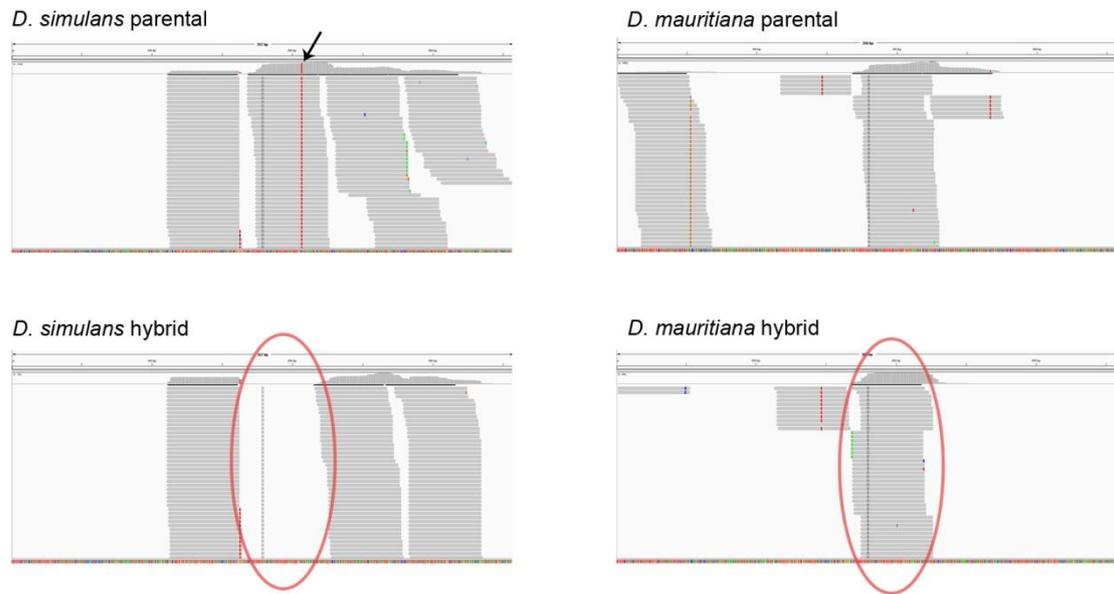


Supplementary Figure 9. Length difference of *D. melanogaster* gene models after reciprocal re-annotation. The annotation of the *D. melanogaster* genome is considered to be the most complete and comprehensive one. After the reciprocal re-annotation of the *D. melanogaster*, *D. simulans* and *D. mauritiana* genomes the *D. melanogaster* gene models could be artificially truncated. This plot depicts the number of gene models that have X% of the original length after the reciprocal re-annotation.

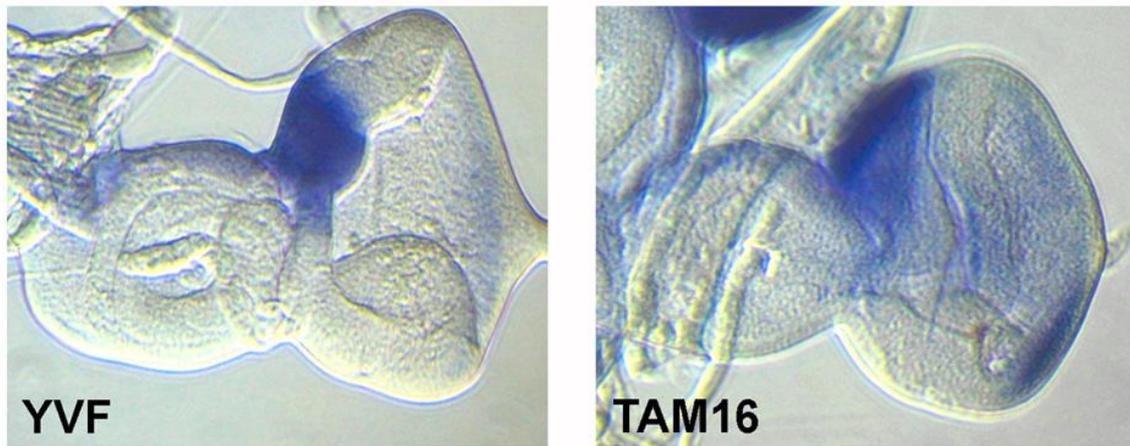


Supplementary Figure 10. Gene expression of conserved genes. Heat map representing the expression of the 1,000 that are most differentially expressed between stages (72h vs 120h) but that have consistent expression across species. Each row in the heat map represents one gene and the color in each cell (from white to dark blue) represents the normalized expression level as indicated in the color key (top left corner). Genes are order by hierarchical clustering based on the distances dendrogram (left side) and grouped into 8 clusters according to their expression profile (different vertical colored bars between the dendrogram and the heat map). The samples are also ordered using the shortest distance method (dendrogram on the top indicates the grouping of the samples, indicated at the bottom of the heat map). On the right side of the heat map are the enriched transcription factors for each cluster of genes as predicted by *i-cisTarget*. The NES score is indicated in brackets for each transcription factor.

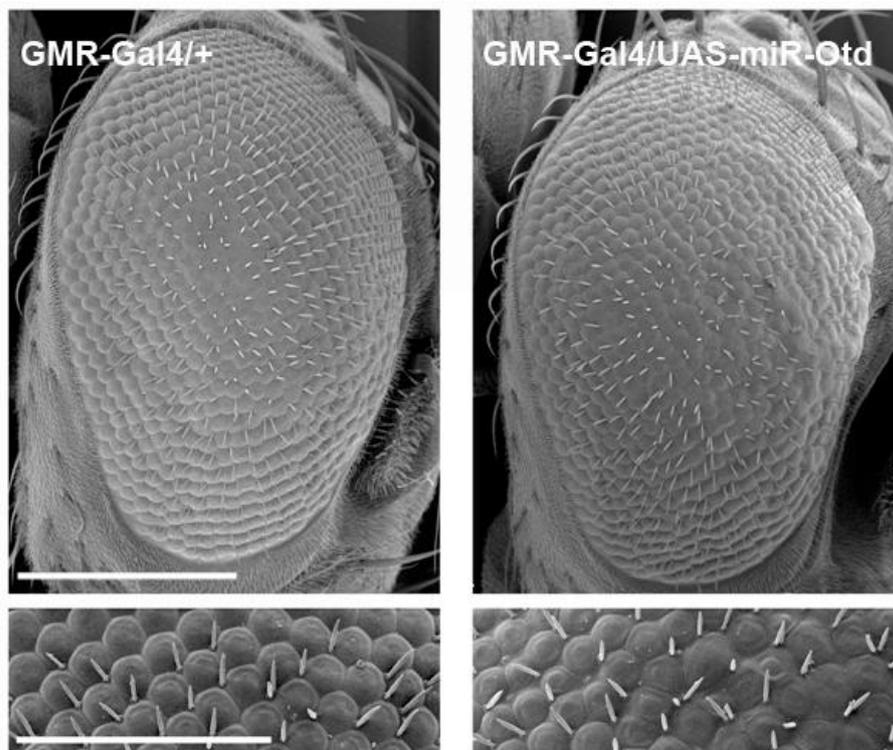
mt:ND3



Supplementary Figure 11. Parental and hybrid reads mapped to mt:ND3. The *D. simulans* reference for mt:ND3 contains a C residue at position 207 bp, while all the RNA-seq have a T (arrow). In *D. mauritiana* there is a T in this position and also its RNA-seq reads have a T. Therefore the hybrid reads map to the *D. mauritiana* allele (red circle), although these reads come from the *D. simulans* allele, and in the rest of the body of the gene the reads correctly map to *D. simulans* variant.



Supplementary Figure 12. *In situ* staining of *ocelliless* in *D. simulans* YVF and *D. mauritiana* TAM16. Image kindly provided by Dr. Isabel Almudi, at the time working at Oxford Brookes University (Oxford, UK).



Supplementary Figure 13. SEM image of *GMR>>oc_{dsRNA}* eye. Image kindly provided by Dr. Isabel Almudi, at the time working at Oxford Brookes University (Oxford, UK).

7.3 Supplementary Tables

Supplementary Table 1. Hb targets in clusters 4 and 11

Gene ID	Gene Symbol	Cluster	Location
FBgn0041171	ago	11	ubiq. lig. compl.
FBgn0039908	Asator	11	cytosol
FBgn0032629	beat-IIIc	11	extracel. region
FBgn0024250	brk	11	nucleus
FBgn0015609	CadN	4	membrane
FBgn0026144	CBP	4	cytosol
FBgn0028509	CenG1A	11	membrane
FBgn0028953	CG14478	11	
FBgn0031632	CG15628	4	
FBgn0029804	CG3097	4	
FBgn0085400	CG34371	11	
FBgn0250867	CG42238	11	
FBgn0259735	CG42389	11	membrane
FBgn0259823	CG42404	4	
FBgn0263392	CG43444	11	
FBgn0264502	CG43901	11	
FBgn0029834	CG5937	4	trans-golgi netw.
FBgn0038676	CG6026	11	
FBgn0032399	CG6785	11	nucleus
FBgn0004396	CrebA	11	nucleus
FBgn0004198	ct	11	nucleus
FBgn0000439	Dfd	11	nucleus
FBgn0042650	disco-r	11	membrane
FBgn0000463	Dl	11	cell surface
FBgn0038071	Dtg	4	nucleus
FBgn0002629	E(spl)m4-BFM	11	nucleus
FBgn0002633	E(spl)m7-HLH	11	nucleus
FBgn0260400	elav	11	cytosol/membrane
FBgn0000635	Fas2	11	membrane
FBgn0011592	fra	11	microtubule
FBgn0259108	futsch	4	membrane
FBgn0001085	fz	11	membrane
FBgn0027343	fz3	11	membrane
FBgn0264574	Glut1	4	nucleus
FBgn0016660	H15	4	
FBgn0035160	hng3	4	lysos. membrane
FBgn0034261	HPS4	11	cytoplasm
FBgn0001226	Hsp27	4	nucleus
FBgn0001235	hth	11	nucleus
FBgn0001269	inv	11	nucleus

FBgn0053182	Kdm4B	11	membrane
FBgn0017590	klg	4	nucleus
FBgn0001320	kni	11	nucleus
FBgn0015721	ktub	11	nucleus
FBgn0026411	Lim1	11	cytosol
FBgn0053087	LRP1	11	nucleus
FBgn0040765	luna	11	nucleus
FBgn0002643	mam	11	nucleus
FBgn0261963	mid	11	cytosol/membrane
FBgn0002932	neur	11	cytosol/nucleus
FBgn0002945	nkd	11	nucleus
FBgn0005771	noc	11	membrane
FBgn0032123	Oatp30B	11	nucleus
FBgn0004102	oc	11	cytosol/membrane
FBgn0020386	Pdk1	11	nucleus/ubiq. lig. comp.
FBgn0013725	phyl	11	
FBgn0264817	pre-lola-G	11	nucleus/cortex/ membrane
FBgn0004595	pros	4	membrane
FBgn0004635	rho	4	
FBgn0031118	RhoGAP19D	11	intracellular
FBgn0083940	RhoU	11	membrane
FBgn0041097	robo3	4	nucleus
FBgn0003300	run	4	membrane
FBgn0003310	S	11	cytosol
FBgn0041094	scyl	11	membrane
FBgn0010415	Sdc	11	nucleoplasm
FBgn0003435	sm	4	nucleus
FBgn0042630	Sox21b	11	cytosol/membrane
FBgn0262733	Src64B	11	cytosol
FBgn0266521	stai	11	membrane
FBgn0020248	stet	4	membrane
FBgn0014388	sty	11	
FBgn0003716	tkv	11	cytosol/membrane
FBgn0026160	tna	11	nuclear chromatin
FBgn0010452	trn	11	membrane
FBgn0004360	Wnt2	4	membrane
FBgn0004607	zfh2	11	nucleus

Supplementary Table 2. Raw values for the length differences of gene models between species.

Length difference	<i>D. melanogaster</i> / <i>D. mauritiana</i>	<i>D.melanogaster</i> / <i>D. simulans</i>	<i>D. mauritiana</i> / <i>D. simulans</i>
Published annotation			
0 bp	6,118	6,228	6,976
1 – 9 bp	2,043	1,773	1,527
10 – 49 bp	1,127	1,081	734
50 – 99 bp	239	495	378
100 – 499 bp	341	376	265
500 – 999 bp	71	27	62
1.000 - 9.999 bp	55	12	50
> 10.000 bp	0	2	2
Total	9,994	9,994	9,994
≤ 49 bp	9,288	9,082	9,237
Direct re-annotation			
0 bp	7,822	7,761	9,847
1 – 9 bp	3,414	3,456	2,046
10 – 49 bp	1,386	1,371	777
50 – 99 bp	143	142	123
100 – 499 bp	262	280	294
500 – 999 bp	147	151	125
1.000 - 9.999 bp	148	160	115
> 10.000	6	7	1
Total	13,328	13,328	13,328
≤ 49 bp	12,622	12,588	12,670
Reciprocal re-annotation			
0 bp	8,811	8,792	10,573
1 – 9 bp	3,238	3,368	2,190
10 – 49 bp	1,191	1,110	522
50 – 99 bp	28	17	12
100 – 499 bp	32	17	10
500 – 999 bp	6	3	3
1.000 - 9.999 bp	5	4	1
> 10.000	0	0	0
Total	13,311	13,311	13,311
≤ 49 bp	13,240	13,270	13,285

Supplementary Table 3. Mapping percentage of *D. melanogaster* replicate A to different references.

Reference	aligned 0 times	aligned 1 time	aligned >1 times	overall alignment
longest coding sequences	40.45%	56.71%	2.84%	59.55%
all coding sequences	40.03%	18.84%	41.13%	59.97%
longest full transcripts	14.24%	80.14%	5.61%	85.76%
all full transcripts	12.62%	25.16%	62.22%	87.38%
genome	4.79%	86.97%	8.24%	95.21%

Supplementary Table 4. Sequences of the primers used for the qPCR experiment.

Gene	primer	sequence	Temperature [°C]	Amplicon size [nucl]
lace	forward	GCACCCGCGTACACTGAAAT	59	91
	reverse	CCGGATGGTAGTTGATCGAGC		
CG3558	forward	ACCTCTTTTCTTCTCCGCC	59	94
	reverse	ATGAAGTTGGTAGTGGTTCGGC		
dac	forward	GAAGCATCGCCTGGACAACG	59	100
	reverse	GATGGGCGGCGGATTGTAG		
RAF2	forward	CAGGCTGCCCAATCTTTACTTCA	57	89
	reverse	TCAGGCCGTCAAATAGCTGT		
Cp110	forward	GAGATGGGAGGTAGCCACAG	61	104
	reverse	GGGTCCATGGAAGTAGAGCC		
CBP	forward	TCGGATGATGAGTTCGAGCG	57	88
	reverse	GCATTTTGCGGCGCCAGAA		
CG6766	forward	TCCCACGAAGCCAAAGATTTT	57	108
	reverse	GTAGAATGTTTCGAGCTGATTG		
CG13784	forward	TTGTCGGTACTTTGGTTTATGGA	57	82
	reverse	GCAGAAATTGTGGCGCCCC		
piwi	forward	TTGGAATTAGTTGGCCGTAATCT	57	98
	reverse	CGAATCGATGTCTCATAGCCCG		
alm	forward	ATTGCAGTGCGGGAATTIGC	59	100
	reverse	AGAAAGGTGAGCATGGTGGT		
Nplp1	forward	TGTGAGTGCTACTGATGATGTCG	60	98
	reverse	CGTGAAGCTGGTACTCGGG		
actin79B	forward	GCCAACCGCGAGAAGATGAC	61	95
	reverse	GAGGCGTACAGGGAGAGCA		

Supplementary Table 5. Differential expression of genes in QTL

Gene ID	Gene Name	%ident	#SNP	Mean TAM	Mean YVF	DESeq Log2FC	DESeq p-adj	edgeR Log2FC	edgeR FDR
FBgn0004102	oc	99.7	1	3440.2	2565.5	-4.23E-01	3.41E-02	-4.25E-01	1.85E-01
FBgn0004656	fs(1)h	62.5	710	51798.7	48313.6	-1.00E-01	7.28E-01	-1.03E-01	8.58E-01
FBgn0004657	mys	98.6	10	38834.0	36864.5	-7.51E-02	7.02E-01	-7.76E-02	9.00E-01
FBgn0010329	Tbh	99.5	3	434.5	70.0	-2.63E+00	4.89E-07	-2.63E+00	2.77E-09
FBgn0011586	e(r)	100	0	5392.7	5696.2	7.90E-02	7.87E-01	7.60E-02	8.72E-01
FBgn0011661	Moe	100	0	130384.8	120548.0	-1.13E-01	4.32E-01	-1.16E-01	8.30E-01
FBgn0014032	Sptr	98.9	3	5231.6	7322.8	4.85E-01	5.76E-04	4.82E-01	1.89E-02
FBgn0017566	ND75	100	0	20665.7	19456.5	-8.70E-02	6.36E-01	-8.96E-02	8.68E-01
FBgn0020653	Trxr-1	99.3	4	55544.5	58217.1	6.78E-02	6.57E-01	6.52E-02	9.16E-01
FBgn0021767	org-1	97.2	20	459.5	624.3	4.42E-01	3.22E-01	4.40E-01	4.77E-01
FBgn0023506	Es2	98.8	6	4377.7	6350.4	5.37E-01	2.99E-04	5.34E-01	7.19E-03
FBgn0025800	Smox	100	0	15568.4	16044.6	4.35E-02	8.15E-01	4.08E-02	9.51E-01
FBgn0025864	Crag	99.8	3	31348.2	33993.0	1.17E-01	6.01E-01	1.14E-01	8.36E-01
FBgn0026318	Traf6	99.6	2	6336.3	6070.0	-6.20E-02	8.23E-01	-6.45E-02	8.98E-01
FBgn0026411	Lim1	100	0	3503.2	3259.1	-1.04E-01	6.75E-01	-1.07E-01	8.47E-01
FBgn0026679	IntS4	99.6	4	10117.2	11463.6	1.80E-01	2.32E-01	1.78E-01	5.35E-01
FBgn0027330	l(1)G0020	99.6	4	15824.0	16482.3	5.88E-02	7.50E-01	5.62E-02	9.21E-01
FBgn0027864	Ogg1	98.3	6	2079.8	2132.4	3.60E-02	9.23E-01	3.34E-02	9.73E-01
FBgn0029992	Upf2	99.2	10	13257.7	11640.4	-1.88E-01	2.29E-01	-1.90E-01	4.89E-01
FBgn0029994	CG2254	100	0	2372.7	400.4	-2.57E+00	4.40E-08	-2.57E+00	1.44E-13
FBgn0029996	Ubc-E2H	100	0	8177.7	8513.2	5.80E-02	7.83E-01	5.54E-02	9.07E-01
FBgn0029997	CG2258	98.5	12	4174.9	3708.3	-1.71E-01	4.92E-01	-1.73E-01	6.86E-01
FBgn0029999	CG1575	99.1	6	5829.4	7592.6	3.81E-01	1.21E-02	3.79E-01	8.12E-02
FBgn0030000	CG2260	99.2	5	7493.1	8539.5	1.89E-01	2.78E-01	1.86E-01	5.34E-01
FBgn0030001	cyr	98.6	7	223.7	357.6	6.76E-01	2.50E-01	6.73E-01	2.20E-01
FBgn0030003	CG2116	98.8	7	7310.6	8225.6	1.70E-01	3.39E-01	1.67E-01	5.86E-01
FBgn0030004	CG10958	99.7	2	7756.3	3828.2	-1.02E+00	1.33E-14	-1.02E+00	2.28E-10
FBgn0030005	CG2120	97.5	8	657.9	615.8	-9.54E-02	8.34E-01	-9.88E-02	9.34E-01
FBgn0030006	CG17982	96.3	11	4376.4	4226.2	-5.04E-02	8.27E-01	-5.32E-02	9.32E-01
FBgn0030007	alpha-PheRS	99.8	1	15497.3	16036.9	4.94E-02	8.38E-01	4.66E-02	9.36E-01
FBgn0030008	CG2129	99.4	3	2537.3	2613.3	4.26E-02	9.21E-01	3.98E-02	9.65E-01
FBgn0030010	CG10959	98.4	7	1420.0	1410.2	-9.93E-03	9.91E-01	-1.26E-02	9.93E-01
FBgn0030011	Gbeta5	100	0	1113.6	1134.1	2.63E-02	9.97E-01	2.30E-02	9.88E-01
FBgn0030012	CG18262	98.1	9	3947.9	4171.4	7.94E-02	7.67E-01	7.66E-02	8.99E-01
FBgn0030013	GIIspla2	98.6	3	2201.6	1719.0	-3.57E-01	1.21E-01	-3.60E-01	4.24E-01
FBgn0030017	CG2278	98.5	10	1478.4	1577.9	9.40E-02	7.76E-01	9.17E-02	9.31E-01
FBgn0030018	slpr	99.4	7	12367.2	11516.3	-1.03E-01	6.36E-01	-1.05E-01	7.91E-01
FBgn0030025	CG2147	98.1	3	3579.0	3689.5	4.39E-02	9.80E-01	4.07E-02	9.59E-01
FBgn0030026	sni	100	0	739.5	1235.4	7.40E-01	1.30E-02	7.37E-01	9.67E-02
FBgn0030027	CG1632	99.7	3	5954.1	3840.4	-6.33E-01	2.11E-05	-6.35E-01	1.70E-03
FBgn0030028	Corp	99.5	1	152.7	149.4	-3.17E-02	9.90E-01	-3.35E-02	9.88E-01
FBgn0030029	CG15343	96.2	8	463.5	1062.8	1.20E+00	4.47E-01	1.19E+00	7.34E-02

FBgn0030030	CG1636	99.7	1	3091.9	2778.1	-1.54E-01	5.20E-01	-1.57E-01	7.65E-01
FBgn0030034	CG10555	33.9	447	16212.6	16046.3	-1.49E-02	9.87E-01	-1.75E-02	9.81E-01
FBgn0030035	CG11190	98.6	9	11345.3	10108.8	-1.66E-01	3.04E-01	-1.69E-01	5.72E-01
FBgn0030037	CG12125	100	0	4818.2	4938.7	3.57E-02	8.67E-01	3.31E-02	9.56E-01
FBgn0030038	CG1440	99.8	1	21372.9	22430.6	6.97E-02	6.98E-01	6.70E-02	9.06E-01
FBgn0030039	CG12123	97.7	5	1921.1	2313.8	2.68E-01	2.85E-01	2.65E-01	6.00E-01
FBgn0030040	CG15347	98.7	3	210.9	298.3	5.00E-01	5.22E-01	4.96E-01	4.89E-01
FBgn0030048	CG12112	99.5	1	2092.6	2733.4	3.85E-01	7.63E-02	3.82E-01	3.04E-01
FBgn0030049	Trf4-1	99.4	6	16358.1	16069.2	-2.57E-02	9.29E-01	-2.83E-02	9.66E-01
FBgn0030051	spirit	96.9	12	3695.9	2202.8	-7.47E-01	9.77E-06	-7.49E-01	2.89E-03
FBgn0030052	CG12065	100	0	6635.0	6191.5	-9.98E-02	7.20E-01	-1.02E-01	8.21E-01
FBgn0030053	CG12081	99.8	1	7215.0	7674.7	8.91E-02	6.52E-01	8.65E-02	8.36E-01
FBgn0030054	Caf1-180	98	24	12724.7	14432.6	1.82E-01	2.00E-01	1.79E-01	5.81E-01
FBgn0030055	CG12772	99.5	4	3923.8	3530.5	-1.52E-01	5.44E-01	-1.55E-01	7.30E-01
FBgn0030056	CG11284	100	0	5893.5	5913.6	4.92E-03	9.93E-01	2.19E-03	9.97E-01
FBgn0030057	Ppt1	98.1	6	5868.1	4683.2	-3.25E-01	3.85E-02	-3.28E-01	1.42E-01
FBgn0030060	CG2004	99.5	2	9997.7	12238.3	2.92E-01	3.50E-02	2.89E-01	2.09E-01
FBgn0030061	CG1785	99.2	4	16725.2	18873.6	1.74E-01	2.23E-01	1.72E-01	6.16E-01
FBgn0030063	CG1789	96.6	8	5274.3	5969.7	1.79E-01	5.47E-01	1.76E-01	6.25E-01
FBgn0030065	CG12075	95.4	46	12867.9	14201.1	1.42E-01	3.20E-01	1.40E-01	7.08E-01
FBgn0030066	CG1885	95.6	11	2953.3	3896.3	4.00E-01	2.61E-02	3.97E-01	1.71E-01
FBgn0030067	Rbm13	97.7	8	9496.0	9200.0	-4.57E-02	7.79E-01	-4.86E-02	9.34E-01
FBgn0030073	CG10962	99.2	2	212.6	211.1	-9.83E-03	9.73E-01	-1.35E-02	9.96E-01
FBgn0040319	Gclc	99	7	12141.5	4429.6	-1.45E+00	8.63E-05	-1.46E+00	1.04E-10
FBgn0040928	CR15345*	60.4	42	245.9	164.3	-5.82E-01	4.37E-01	-5.85E-01	3.33E-01
FBgn0040929	CG12659	100	0	1222.8	1051.9	-2.17E-01	5.18E-01	-2.20E-01	7.78E-01
FBgn0041629	Hexo2	99.5	3	5933.7	7891.9	4.11E-01	6.05E-01	4.09E-01	3.21E-01
FBgn0053181	CG33181	99.6	3	2698.7	3880.9	5.24E-01	1.34E-01	5.22E-01	1.01E-01
FBgn0259734	Nost	98.9	15	2575.9	2509.4	-3.77E-02	9.57E-01	-4.03E-02	9.66E-01
FBgn0261549	rdgA	99.2	12	5321.7	4069.4	-3.87E-01	2.36E-01	-3.89E-01	2.21E-01
FBgn0261793	Trf2	90.4	82	10699.0	10942.3	3.24E-02	8.32E-01	2.99E-02	9.62E-01
FBgn0261873	sdt	98.5	31	29017.8	26778.0	-1.16E-01	6.60E-01	-1.19E-01	8.16E-01
FBgn0262976	lawc	98.6	1	748.9	906.8	2.76E-01	4.86E-01	2.73E-01	7.38E-01
FBgn0264975	Nrg	93.6	84	107235.6	112760.2	7.25E-02	5.89E-01	6.98E-02	9.16E-01

* pseudogene

7.4 Sequences of cloned QTL candidates

>Dsim CG1885

ATGACGAGTCGCCAGCGAACTGTGATCATATTCAAATCGGAGTCGGAAAAGCAGCGATGTGTA
 CGCGGAAACGCTGGAGAAGCACGATTTCAATCCTGTCTTCGTGCCACACTGAGCTTTGGCTT
 CAAGAATCTGGAGGAGCTGCGCGCCAAGCTCCAGAATCCGGACAAGTATGCCGGCATCATAT
TCACATCGCCGCGCTGCGTGGAGGCGGTGGCTGAATCCCTCAATCTCGGCGAGCTGCCCGGC
 GGTTGGAAGATGTTGCATAACTATGCCGTGCGGCGAGGTGACCCACAATCTGGCGCTGAGCAC
 CTGGACCAGCTATTCACCCACGGCAAACAGACGGGCAATGCCCGGGCACTGGGCGACTACA
 TAGTGGACACGTTTCGATGGATCGCGCGCCCTGCTGCTGCTGCTGCCGTGCGGCAATCTGGCC
 ACCGATACGCTGCTCTCAAGCTGGCCGAGAATGGCTTCTCCGTGGACGCGTGCGAGGTGTA
 CGAGACGCGCTGCCATCCCCGAAGTGGGCGCAATGTGGAGCGGGCACTGGAGATCTACGGC
 GAGTCGATCGAATTCCTTGCCTTCTTCTCGCCGTGCGGCGTCAATTGTGCGCAGCAGTACTTC
 ACCGCCCGCAGAATCGCTGGGCCAGAAAGGTCTACTGCACCCGAGCGGCCGACGGGTGGA
GCATCTGGTCAAGGTGCTGCTCAATCCGCAGGACAGCCGGGAGCGGCTGCTCAAGGAGC

>Dsim CG10958

ATGGACGACAACGAGGATGAACTGGAGGAGCAGCAGGAGATAGTCAGCGACGGCAGTGTG
 GAGGAAGAGGAGGAGGTGGAGCCTGATCTGGGACCAGTGGACAGTTGGGAGTCTTTAAT
 GACCGCATCGATGGCCTGATCTTCAACGAGCGATGCGACAAGTCGGTGAAGAAGTTGGACG
 AACGCCGGCTGGCGATACTCAACCGGGTTCGTGAGCAGTTCGCGAGGCCCGGAAAGGCGTC
 GCCGAACGTTGCCAGGCCAGAAAGTCGCCATTGACCAACGGCTGGAGTTGTCCAGCGAGCG
 ATTGAACGAACTGGTACGCTTTGGCAAAGAGCTGGTGACCAATGTCCGGGTGGCCAACGAG
 CGAAGGGAGCTGAATCGCCGAATCTTCGAGGGTGCAGCAGAAGAATCAGATGAATGTGAAGC
 TGCAACCGGAAAGCGTTGAGACAATGGCTCGTTTCGAGAACATTAAGGCGCGCTGGACGGA
 GCTGGAGGAGACCAACGAGCCGATGCTGCTGTGGGACCAAATCGAAGAGCAAAAAGAAACGC
 ATTGCCGAGATCATGGCACGCAAGGATGAGATGATATCTGCCTGCCAAGCGGAGGTGGACC
 GCATGAATGCCAAGTACGAGTTCGATCGCGAGCGGAGGCCAGGATCTTTGCTGTCTGGTG
 GAGCGCGTCCGATCACCAGGTGGAGACGCTAAAGGAGGCCCTACAAGGAGCACATCCAAATGC
 TACGCCAAACCATCGAGGAGGAGCGGCAAAATCTTCGCCGATAATGCGGTGGAGAAGTGCCG
CACCTTCTTCGATGCGATGAATGCCAACTTTGACGAGAAGGCCAATTTGGTAAGGGCACGCG
 AGCAATTCTATGCTCGCCAAACCCAGCAGATCAACGAGTCCAGGAGGAGCTGACCAAGAGC
 ACTCGCATTCGGCTGGAGAAGGAGTGCAGAGGCTCGAGCTGGAGCTGCGTCGCACGCGCG
 ACAATGTGCTGATGAACTCCGAGAAGCTGGACTATAACTATCAAGTTCTCCAGAAGCGCAAC
 GAGGAGAACGTGATCATTAACAACCAGCAGAAGCGACGGGTGGCCCGACTGCATGAGGCCGA
 TCGGACGCACCCGGCGTGGCCTGAAGAATCTTTACAACACGGGCAAGCGGAACATAGCCCGT
 CTCTCCTCGGACATCTACAAGCTGCACTCCAACATCAACGATATGGAGTCGAAGGCGCATCAG
 GCGAGGCTGAACAATCGTGAGAAGTTCGATCGCATCTGGGAGATCAACTACAAGGAACTGAA
 CCTGCTGGTGGATCGGGTCTACCACATCGATCGCATCATAACGAGCAGCAGCTGGCCATGC
 CGTGGTCCAGTCCCCTGCCACCTATTCCGAACATCAACAAGGCGAAGAAGAAGCGCAATAAC
 ATCTGGAGAAGTTTGACATGCGGATTGGCCGGGTTCCCAAGAAGGGGTGATGCCCAAGTC
 GAGCGTAAATTCGAAGCCGGATATCAAGGAGCTGCCGCCGATTTCGCTACGCTTGATGCGTA
 ATCTCATTTCGAAGCTGTCCGATCGCGCGGTTCCTCATCGAGGAGCGCCTGCTAAAGATTC
 TTGAGCCGTATTCCGAGGAGGAAAAATGCCTGGTACGCATCGACAATATATTTGCGGCGCTG
 CGAATTCGTCATCTGCGTGACGTCAAGGAAGTACCAAGGTTTTCATGCCGTACACCTACTGT
 CCGAATTGCCAACCGCAGGGATTGAGTCCACGCAAGTGCGCCGAGGTTCATGAAGGATCA
 AAAGCCAAATCGGCTGCAGGGCACGGCGAGTGGCAAACCGGAGGAGAAGGAGCATAGGTC
 GACAGGGGAAACATTTCTACCCAAGAGCGACGAGGCGGCCAAGAGATGCCACAATCACTATC
 TGGTCAATGGAGCCGGCCCTCTGCCTGCACGCCATGAATCTGTTACCTCCAAGATGCACAAAA
 AGATGTACGAACACGAACCGGGCAGCATTCTCAATGCGGTAATCTTATTACAGATTACGGAT
 GCAGAGATCCGTAATTTCTGGCGCCAGTTCACGCTGCTTCCCAGGCTCCAAGTGAAGCTG
 TGGAAGACCCTGGAGCACGGCCTGAATCACTACGTGGAGGTGCTCAAGATGCGTGTGCAGT
 ACGATGCGGAAGTCGTCTTCTTGCCTGCCAGAACGAAGAGTTGCGCCATCTGCTGCAGAAG
 TTCACCGTC

>Dsim CG1632

ATGTCGGACATATACTACTGCAAGCAACAGCCAGATGAAGAAATCGCGCGACGCGGATAAATC
 ATCCGCCCGGTTGTCCGCGGCCAATATCAATGCGAATGCGACCAATTTGTCCGGCTGCTGATAA
 GAAGAACAACAACAACAACTGCAGCAATCGCAACAAAGAGAAGAGCATGCAGGCGAACGGT

AAAAATTGGAAGGGCATTGTCCAGAGTAATCCCAATCCGAGTGGCGGTGTGGGCACCACGG
 CTCAGCCGCCAAAGTTTTCCACAACACCCGATGCACAAGGCACCACAAACCGCACCACAGTC
 ACGGTAATGGAAGTGCAGCTTCCGCCCGGTAGGAGCAGCAGGCCCTGGCCAATGGACAACC
 ACCACTGGAAGCCACACAGCAGCGGGAAAGGGAAACGGGAACGAGATCGGGAGCGGGAAACG
 GGAAAGAGACAGGGAGCGAGACAGAGCGGGAAACACCAGCTTCATATGCACCAGCACAAAT
 CACGGGCTGCGAAGAAAATCCGAGTCCGTCTGTCCACCGACTCGGACATCCGCTTCACCCGC
 CGGAAACTGGGGCAGTGGTCAAAAAGTGGCGTGTGCCGTCATCGCTGGATTCTCATCGCCCT
 CCTCGTCGGGAATATTTGTCTATGTGGGATATACCTATTTCCGACCGGAGCCGCTGCCAGA
 TCGCGTTTTCCGCGGCCGCTTCATGGTGTGAACGACAAGTGGAGCATGGAGCTGGCCAACC
 AGAACTCGATGAGGTTCCAGCACAAGGCGCGACTACCGGGAGCGGATCAATCTCACCCCTG
 CGCCGATCCGATCTGCGGGAGGCCTACGAGGGCAGCGAGATTTTGGCCCTGGATGGGAGCG
 AGGATAACAACAACATAGTCGTTCACTTCAACATGATCTTCGATCCGTACGCGGGTCTGGTGA
 GCAGTGGTGACCTTTTGGCCCTATTCACAGGAGATGACCCAGCCGCCGAGCAGCGCCGC
 CATTCGCCAACATGACGGTGGATGTGGCCAGTTTGAGCATCAAGGAGACGACCCGGCCTGAT
 CGAGGAGCCCCTGATGTCCAGTTCGCCGCTGGGCGGACACGATGAGACCACCGAGCCGGTG
 GTGACCACCACTCCGGCTCCGCCACGTGCTGCTCACCCTGGAGCTATCCTACTGCCGCCAG
 GTGGGCTACAACATTACCACCTATCCGAATCTTCTGGGACACGCCAGTTACGAACAGTTGGCC
 GAGGACGTGATCGTGTTCGGGAACTGGTGGACGGTGAATGCCATCGGGAGGCCTACGACT
 TTGTGTGCCGGCTCCTCCAGCCGCCGTGCGACACGCACGGCTCCGATATGCAGCCAACTCCG
 GGCCAGATATGCCGCGAGTACTGTGAATCCTTCATGGCCGGTTGTGGCGGTGATGCGCCGA
 GCGCTTCGGCAGTTTTCGACTGCGAACGCTTCCCGGAGTCCACGGGCACCCAGTCGTGCC
ACCAGAAGCCGCACTGTGTCAGCGACATGCAATCCAATGTCCAGAGTCCCTCGGCTCTGCGAT
 GGCTATGCCGATTGTCCGGATCTTCCGACGAGCGCAGCTGCGCCTTCTGCTCGCCCAACGCT
 CTGTATTGTGGCGTGGCAGGGCGTGTGTGCCCGCAAGGCACGATGTGATGGCAAGCGC
 GACTGTCCGGACGCGCCGATGAGAAGGATTGCCATCTATAGCTCCACTGCGCCCGGATCT
 GCTGCAGCCGGAGCCCTGGTACCCTACCTCTCCCGCTTCCATTCCGCCGGCTACGCCGTCTT
 CTCCGAGAAGGGAGTGGTGGCAAGCTGTGCGCCGAGGGTCTGGAGGGCGATGCCAAGCT
 GGTGGTGCGCCAAACGGTCTCCGAGTCGCTTTGCAAGTCCCTGGGATACGAATCCGTGGAAA
 TATTCGACGTGCAGAACGATACGGAGCGTTTGAACGACTACGTGCGTGTTTGGATCCACAT
 GCGCCGAGATCAGCTTCATACGGACGCACTGCCCGGCGACAAGTGTATACGTGGGCTG
 CGGGGAGCTTCGCTGCGGCGTCCAGTCCGCGCTTTTCAATGCCAAGCAGCACCTCTCGCTGC
 CGAAGATGTCCGCTCCTGGGGATTGGCCCTGGCTGGTGGCCCTGTTCCGCGAGGATATCCAC
 GTTTGCGACGGCACCTGATCTCGCAGGACTGGGTCCCTACCACCGAGGGCTGTTCCAGGG
 CCAGCCGCGTGCCACTTGGATGGCCATTGTGGGCGCAGTTCGTCTGTCCGCCAAGGCACCGT
 GGACGCAGAGGCGCCGCATCATTGGCATGATCAAGAGCCCGGTAGAAGGTTTCGACGGCGGC
 ACTGGTGCCTGGAGACTCCGGTCAGTACTCGGATCATGTGCGACCCATTTGCCTGCCGG
 ACGCCCTGCAGAGACGCCTGCTCCAGCAACCACCGGCCAGAGGAGATCCCATGTTCCGGTG
 GCCGAGCGATTGGAGGGTCAGCTTGTGAGTCAGCAGCGAAGTCGATTGTGCGAGGAGAACC
 AGCAGTCTTCTCCTATCCCATCGCAGGAGCAGCAGGACTCCACGGAGAATCAGGAGGAT
 GAGGATCAGGACGAGCAGGAGGATCACTTTGGCGGCGAATCAGCCGCTCCTATATGCCCAA
 AGCGGAGGCTTTGCACCAAGAATTGGATGGGTACCCACTGCCGGATCATGCGCCGAGGTTA
 ATTACTACTCCTCCTCCTCCACGGTCACATCCTCCTCCACCGCCGCCGCACAGCCACCAAGGC
 TCCCGTCTTGGCGGCTGTTCGGGCTGCCAGGAGCAGATCTGGACAAACTGCAATACACTCG
 GTTGGTCCCGGCAGCGGGATCACCTGCAGCGTGTCCAGCTCAAGATGGGCGACATGGCGCC
 CTGCGAGAACGTGTCCATTGCCACCGTGAACCTCCATGTGCATGGAGGCCACCTACCAGAAGT
 ACGACTGCACGCAAGAGGAGTATTCGGGAGCGCCCTCCAGTGCCTAATTCGGGGAACGAAT
 CAGTGGGCGCTCATTTGGGGTCTCCTCCTGGCGGATCGCGTGCAGGACCCACGGGCGTGGAGC
 GGCCAGGATGTACGACAAGATCGCCTCGAATGCCGCTGGATCCGCGAGACGATCAGCGC
 GATA

>Dsim_Sptr

ATGGACCTGAAACAGCGCACCTATCTCCTGGTGACCGGGGCATCCCGTGGAATTGGCCGTGA
GTTCGCCCAGCAGCTGGCCAAACGGATCAAAGCCGAGGGTTCCGTGGTGACGCTTCTGGGAC
 GCAATCAAACCCTTTTGCAGGAATCTAAGGCAGAGATTGTGGCCACAGTGCCGGATCTACCC
 GTGCAAACCTACTCGCTGGAGCTGGAACCGCCAAACCGAGGACTTTACCAAGATCTTGGGA
 GGCATCCGCTGGAAAGAACAGTTTTGAGCGAGCCATAGTCATTATAATGCCCGCACAGTGG
 GCGACAGTCCAAGAGGGCCAAGGAAATCGGAGATACGGACTTCTGACGCTACTACTACAC
 TCCAATGTCTTCTCGGCCATTTGCTGAACTGCGAGTTCATGCGCGTCTTCCAGGGAAATCCCA
 AAGTTGGTGGTTAATCTCAGCACCTTGGCAGCCATTGCACCTATATCCTCGATGGCACACTAT
 TGCACGGTGAAGGCTGCCCGTGAGATGTACTTCCGAGTGTGGCCACCGAGGAGTCCGCCG
 AGGACACCTGGTGTGAACTACGCGCCCGGCTCATAGACACGCAGATGACCGTCCAGGTT

CAGCGAGAGGCCACGATCCGGCCGTGGTCGCCATGTTCCGAGAGCAAAGGGAGTCCAAGACCATGCTGACTCCCGCCAGACGACGGAGCGGTTTCATCAAGGTCTGGAGGCATTCAAGTTC AAGTCCGGCGATCATGTGGACTACAGGGATGAGCAGTTC

>Dsim_sni

ATGAACTCCATCCTGATAACCGGCTGCAATCGAGGATTGGGTCTGGGCCTGGTCAAGGCGCTGCTCAATCTTCCCAGCCGCCGACGATCTATTTACCACCTGCCGGAATCGCGAGCAGGCAAA GGAGCTGGAGGATCTGGCCAAGAAGCACTCGAACATCCACATCCTTGAGATTGATTTGAGGAATTTTCGATGCCTATGACAAGCTAGTCGCCGACATCGAGGGCGTGACCAAGGACCAAGGCCTCAATGTGCTCTTCAACAATGCCGGCATAGCGCCAAATCGGCCAGGATAACGGCCGTTTCGATCGCAGGAGCTGCTCGACACCTTGCAGACCAACACGGTGGTGCCATCATGCTGGCCAAGGCGTGTCTGCCGCTCCTGAAGAAGGCAGCCAAAGCGAACGAATCCCAGCCGATGGGCGTGGGCCGTGCCGCCATTATTAACATGTCCTCGATCCTTGGCTCCATCCAGGGCAACACGGACGGCGGAATGTACGCCTATCGCACCTCCAAGTCGGCCTTGAATGCGGCCACCAAGTCGCTGAGCGTGGATCTGTATCCGCAGCGCATCATGTGCGTCAGTCTGCATCCTGGCTGGGTGAAAACCGACATGGGTGGCTCCAGTGCGCCCTTGGACGTGCCACCAGCACGGGCCAAATTTGTGCAGACCATCAGCAAGCTGGGCGAGAAACAGAACGGCGGCTTTGTCAACTACGATGGCACTCCGCTGGCCTGG

8 Curriculum vitae

Personal details

Name	Montserrat Torres Oliva
Nationality	Spanish
Place of birth	Barcelona
Date of birth	02.01.1988
Address	Steinweg 12, 37077 Göttingen, Germany
Email	mtorres@gwdg.de

Education

2006 – 2010	4-year Bachelor of Science in Biotechnology Universitat Autònoma de Barcelona (Spain)
2010 – 2011	Master of Science in Advanced Genetics Universitat Autònoma de Barcelona (Spain) Master Thesis in the Department of Genetics and Evolution Universitat de Barcelona (Spain)
2012 – present	Doctorate in Biology Georg-August-Universität Göttingen (Germany)

Work experience

2009 – 2011	Internship: Bioinformatics support Plataforma Bioinformàtica de la UAB (Spain)
2011 – 2012	Marie Curie Early Stage Researcher: main coordinator of <i>Strigamia maritima</i> genome annotation Department of Zoology, University of Cambridge (United Kingdom)

Scientific meetings

February 2013	First International SpiderWeb Meeting (oral presentation) Oxford (United Kingdom)
March 2013	18 th Evolutionary Biology graduate meeting of the German Zoological Society (poster presentation) Göttingen (Germany)

- September 2013** 23rd European Drosophila Research Conference (poster presentation)
Barcelona (Spain)
- July 2014** Euro Evo Devo Vienna 2014 (poster presentation)
Vienna (Austria)
- November 2014** iSEQ - Methods and applications of Next Generation Sequencing
in evolutionary research (poster presentation)
Leipzig (Germany)
- September 2015** 24th European Drosophila Research Conference (poster presentation)
Heidelberg (Germany)

Attended courses

- July 2014** External Methods Course: EMBO Practical Course on Genotype
to Phenotype Mapping of Complex Traits
Hinxton, Cambridge (United Kingdom)

Language knowledge

- Mother tongue** Spanish and Catalan
- Proficient (C2)** English
- Independent (B2)** German
- Basic (A2)** French and Japanese

Publications (published and submitted)

Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, **Torres-Oliva M**, *et al.* (2014) The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*. *PLoS Biol* 12(11): e1002005. doi:10.1371/journal.pbio.1002005

Torres-Oliva M, Almudi I, McGregor AP and Posnien N. (in revision in *BMC Genomics*) A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species.

Torres-Oliva M, Almeida FC, Sánchez-Gracia A and Rozas J. (in review in *Genome Biology and Evolution*) Comparative genomics uncovers unique gene turnover and evolutionary rates in a gene family involved in the detection of insect cuticular pheromones.