

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

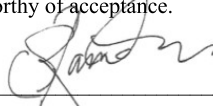


DISSERTATION ACCEPTANCE CERTIFICATE

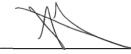
The undersigned, appointed by the
Department of Molecular and Cellular Biology
have examined a dissertation entitled
Computational Approaches to Developmental Biology

presented by Leo Blondel

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature  _____
Typed name: Prof. Sharad Ramanathan

Signature Craig P. Hunter
Craig P. Hunter (Nov 24, 2020 11:46 EST) _____
Typed name: Prof. Craig Hunter

Signature  _____
Typed name: Prof. Allon Klein

Signature _____
Typed name: Prof.

Signature _____
Typed name: Prof.

Date: November 23, 2020

Computational Approaches to Developmental Biology

A DISSERTATION PRESENTED

BY

LEO BLONDEL

TO

THE DEPARTMENT OF MOLECULAR AND CELLULAR BIOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE SUBJECT OF

COMPUTATIONAL BIOLOGY

HARVARD UNIVERSITY

CAMBRIDGE, MASSACHUSETTS

NOVEMBER 2020

©2021 – LEO BLONDEL
ALL RIGHTS RESERVED.

Computational Approaches to Developmental Biology

ABSTRACT

The origin and evolution of new genes is an active topic of research, relying on the taxonomical diversity now present in sequence databases. Using those databases, we described how *oskar*, a key determinant of germ cell determination, likely arose from a horizontal gene transfer and then described its evolution and conservation in insects. The number of ovarioles, the egg-producing unit of the insect ovary, is hypothesized to inform the individual's reproductive capacity. Using network biology approaches, we analyzed the effect of signaling pathway genes on the number of ovarioles and eggs laid by *Drosophila melanogaster*. We found putative gene modules regulating both traits and predicted novel genes affecting both phenotypes. The specification of germ layers is a central mechanism of the embryogenesis of animals, but the underlying molecular mechanisms have only been extensively studied in model organisms. Using *Parhyale hawaiiensis*, a crustacean amphipod, I generated preliminary methods for the generation of single cell RNA sequencing of early embryogenesis, as well as recorded with light sheet microscopy the first three days of embryogenesis. The preliminary analyses of the sequencing datasets were inconclusive, but, analyzing one of the microscopy datasets, I described new preliminary cellular dynamic results. Finally, to observe and annotate 4D microscopy datasets, I developed a tool that allows the visualization of large volumetric datasets in Virtual Reality.

Contents

0	INTRODUCTION	1
0.1	The origin and evolution of <i>oskar</i>	3
0.2	The origin of new genes	3
0.3	The evolution of <i>oskar</i>	6
0.4	The regulation of <i>D. melanogaster</i> ovariole formation and egg-laying	8
0.5	The unique features of crustacean early embryonic development and germ layer specification	10
0.6	The molecular basis of <i>P. hawaiiensis</i> germ layer formation	11
0.7	<i>P. hawaiiensis</i> early embryogenesis	11
0.8	The interactive visualization of microscopy images	13
	References	14
1	BACTERIAL CONTRIBUTION TO THE GENESIS OF THE NOVEL GERM LINE DETERMINANT OSKAR	26
1.1	Introduction	29
1.2	Results	30
1.3	Discussion	39
1.4	Methods	40
1.5	Acknowledgments:	46
	References	46
2	EVOLUTIONARY HISTORY AND FUNCTIONAL INFERENCE OF THE OSKAR PROTEIN	55
2.1	Introduction	58
2.2	Methods	61
2.3	Results	71
2.4	Discussion	95
	References	97
3	TOPOLOGY-DRIVEN PROTEIN-PROTEIN INTERACTION NETWORK ANALYSIS DETECTS GENETIC SUB-NETWORKS REGULATING REPRODUCTIVE CAPACITY	105
3.1	Introduction	108
3.2	Results	110
3.3	Discussion	137
3.4	Methods	142
	References	151
4	STUDYING GERM LAYER SPECIFICATION IN THE EARLY EMBRYO OF <i>Parhyale hawaiiensis</i> WITH SINGLE-CELL RNA SEQUENCING	163
4.1	Introduction	166
4.2	Methods	176
4.3	Results	188
4.4	Discussion	218
4.5	Detailed protocol for <i>P. hawaiiensis</i> embryo single-cell dissociation	220
4.6	List of overrepresented sequences in the two libraries	227
	References	228

5	STUDYING GERM LAYER SPECIFICATION IN THE EARLY EMBRYOGENESIS OF <i>Parhyale hawaiensis</i> WITH MODERN MICROSCOPY	239
5.1	Introduction	242
5.2	Methods	247
5.3	Results	258
5.4	Discussion	273
	References	274
6	DIVING INTO THE THIRD DIMENSION, VIRTUAL REALITY FOR THE ANALYSIS OF VOLUMETRIC MICROSCOPY	281
6.1	Introduction	283
6.2	Methods	292
6.3	Results	295
6.4	Discussion	309
	References	310
7	CONCLUSION	317
7.1	The era of omics	317
7.2	<i>oskar</i> , a novel gene with an intriguing evolution	318
7.3	The regulation of the development of the <i>Drosophila</i> ovary	320
7.4	Towards a better understanding of germ layer specification in <i>P. hawaiensis</i>	321
7.5	Expanding our perception with a third dimension	324
	References	324
	APPENDIX A CHAPTER 1: SUPPLEMENTARY DATA	332
	APPENDIX B CHAPTER 2: SUPPLEMENTARY DATA	353
	References	361
	APPENDIX C CHAPTER 3: SUPPLEMENTARY DATA	369

Listing of figures

1.1	Sequence analysis of the Oskar gene.	33
1.2	Phylogenetic analysis of the LOTUS and OSK domains.	34
1.2	(continued)	35
1.3	Hypothesis for the origin of <i>oskar</i>	38
2.1	Overview of the Oskar protein.	60
2.2	Presentation of the <i>oskar</i> ortholog detection pipeline.	74
2.2	(continued)	75
2.3	Summary of <i>oskar</i> distribution and expression in insects.	76
2.3	(continued)	77
2.4	Phylogenetic reconstruction of hymenopteran Oskar sequences.	80
2.5	Differential conservation of amino acids between hemi and holometabolous sequences.	84
2.5	(continued)	85
2.6	Conservation analysis of the LOTUS domain.	88
2.6	(continued)	89
2.7	Conservation analysis of the OSK domain.	93
2.7	(continued)	94
3.1	Screen methodology.	113
3.2	Relationship between Egg Laying and Ovariole Number phenotypes generated in the screens.	117
3.2	(continued)	118
3.3	Enrichment of genes of individual signalling pathways among the experimentally obtained positive candidates of each screen.	121
3.3	(continued)	122
3.4	Screened genes function as a network.	125
3.5	Representation of Seed Connector Algorithm and output.	129
3.5	(continued)	130
3.6	Phenotypically separable sub-networks formed by analysis of the combined genes from all sub-networks.	132
3.7	Positive prediction rates of the connector genes in each of the four sub-networks.	136
4.1	Life cycle and early embryogenesis of <i>P. hawaiiensis</i>	168
4.1	(continued)	169
4.2	Schematic representation of the blastomere ablation and intra-germ layer compensation experiments for the ectodermal lineages	170
4.3	Schematic representation of the blastomere ablation and intra-germ layer compensation experiments for the mesodermal lineages	171
4.4	Schematic representation of the Maternal to Zygotic Transition of <i>P. hawaiiensis</i>	173
4.5	Schematic representation of the concept of the geo-positioning system	176
4.6	Schematic representation of the creation of a mesh trap for the capture of <i>P. hawaiiensis</i> adults.	178
4.7	Schematic representation of the mounting procedure on a glass slide and coverslip for <i>P. hawaiiensis</i> embryos.	182
4.8	Representative images of heat-shocked embryos and control embryos.	190

4.9	Representative images of <i>P. hawaiiensis</i> embryos stained with antibodies against different phosphorylation states of the RNA Polymerase II CTD.	191
4.10	Positive control for the antibody staining.	192
4.11	Light sheet imaging of <i>P. hawaiiensis</i> embryos stained with antibodies against different phosphorylation states of the RNA Polymerase II CTD.	194
4.12	Bioanalyzer traces of 8 CelSeq2 library attempts.	196
4.13	Proportion of reads per droplet that mapped on the mitochondrial genome and the nuclear genome of <i>P. hawaiiensis</i>	203
4.14	Proportion of reads per droplet that mapped onto a gene annotation against the reads that mapped to the nuclear genome of <i>P. hawaiiensis</i>	204
4.15	Analysis of the number of unique mRNA molecules captured by inDrop.	205
4.16	Preliminary analysis of the pilot scRNAseq data for II S6_7.1.	208
4.16	(continued)	209
4.17	Preliminary analysis of the pilot scRNAseq data for the II S11.1 and II S11.2.	212
4.17	(continued)	213
4.18	Heatmap representation of the expression level for the top 10 marker genes	214
4.19	Comparative analysis of the scRNAseq data for the cells coming from S6, S7 embryos (12-24hpf), and the later stages S11 (64-68hpf) embryos.	215
4.20	Expression level analysis for Ectoderm gene markers collected from the literature.	216
4.21	Schematic representation of the plate with Elmer's paste seals, parafilm, and lab tape.	223
4.22	Schematic representation of the concentration steps.	226
5.1	Example micrographs showing the contrast generated by darkfield microscopy.	243
5.2	Schematic representation of the creation of an agar step mold for injection of <i>P. hawaiiensis</i> embryos.	250
5.3	Custom mounting procedure for multiple <i>P. hawaiiensis</i> embryos on the SimView scope.	252
5.4	Representative images of <i>P. hawaiiensis</i> injected embryos.	259
5.5	Maximum intensity projection of fused light sheet recording of <i>P. hawaiiensis</i> embryos.	260
5.6	Results from the automated segmentation and tracking performed using Ilastik and manual tracking and corrections.	263
5.7	Analysis of cell division rates and nuclear velocities in the tracked WT01_11-17.	266
5.8	Analysis of blastoderm and germ lineage dynamics during the early embryogenesis of <i>P. hawaiiensis</i>	267
5.8	(continued)	268
5.9	Analysis of cellular movement during gastrulation and formation of the midline.	269
5.9	(continued)	270
5.10	Schematic and visual representation of the blastomere ablation experiments.	272
6.1	History of VR headset technological development.	287
6.2	Construction of 2D, 3D images and stereoscopy.	289
6.3	Schematic representation of Ray Marching.	290
6.4	Example images of <i>P. hawaiiensis</i> embryos at the 16 cell stage (stage S5) resulting from the different iterations of the volume rendering shader.	302
6.5	Schematic representation of different 3D image file format strategies.	304
6.6	Blueprints of the VR room user experience.	305
6.7	Examples of the VR user experience shown in the form of snapshots of user views from within the VR.	306
6.7	(continued)	307
A.1	LOTUS Domain RaxML MUSCLE Tree.	333
A.1	(continued)	334
A.2	LOTUS Domain Bayesian MUSCLE Tree.	335
A.2	(continued)	336
A.3	OSK Domain RaxML MUSCLE Tree.	337

A.3	(continued)	338
A.4	OSK Domain Bayesian MUSCLE Tree.	339
A.4	(continued)	340
A.5	SOWHAT constrained trees and results.	341
A.6	LOTUS Domain RaxML PRANK Tree.	342
A.6	(continued)	343
A.7	OSK Domain RaxML PRANK Tree.	344
A.7	(continued)	345
A.8	OSK Tree PRANK Comparison.	346
A.9	LOTUS Tree PRANK Comparison.	346
A.10	LOTUS Domain RaxML T-Coffee Tree.	347
A.10	(continued)	348
A.11	OSK Domain RaxML T-Coffee Tree.	349
A.11	(continued)	350
A.12	OSK Tree T-Coffee Comparison.	351
A.13	LOTUS Tree T-Coffee Comparison.	351
B.1	Summary statistics.	354
B.2	Genome and Transcriptome quality correlation to oskar discovery.	355
B.2	(continued)	356
B.3	Loss of <i>oskar</i> in Lepidopteran.	357
B.3	(continued)	358
B.4	Complete Hymenopteran Oskar phylogeny.	362
B.4	(continued)	363
B.5	Tissue and Stage metadata analysis of Oskar presence in transcriptomes datasets.	364
B.6	Multiple Correspondence Analysis (MCA) of Oskar, OSK and LOTUS.	365
B.7	Evolution of the structure of Oskar in Diptera.	366
B.8	Oskar domains secondary structure conservation.	367
C.1	Violin plots of egg laying and ovariole number of controls in each screen batch.	371
C.1	(continued)	372
C.2	Enrichment/depletion analysis of the 273 signalling pathway genes above the threshold $ Z_{gene} > 1$ (Figure 3.1a) against all signalling candidates.	373
C.3	Comparison of egg laying candidate genes by pathway.	374
C.3	(continued)	375
C.4	Comparisons of the Z_{gene} scores of the positive candidate genes sorted by centrality metrics..	376
C.4	(continued)	377
C.5	Comparison of network metrics of seed lists obtained from the screen.	378
C.5	(continued)	379
C.6	<i>Hpo[RNAi]</i> Egg Laying Sub-Network generated by the Seed Connector Algorithm (SCA).	380
C.7	Egg Laying Sub-Network generated by the Seed Connector Algorithm (SCA).	380
C.8	<i>Hpo[RNAi]</i> Ovariole Number Sub-Network generated by the Seed Connector Algorithm (SCA).	381
C.9	Centrality metrics of the sub-networks.	382
C.10	Comparison of network metrics after application of the seed connector algorithm (SCA).	383
C.11	Signalling pathway enrichment/depletion analysis.	384
C.12	Comparison of edge densities between the seven sub-networks of the meta network, and to a randomly assigned grouping of genes in the meta network.	385

TO MY DAD, VINCENT BLONDEL, WHO LEFT US MUCH TOO EARLY. YOU PUSHED ME TO OVERCOME DIFFICULTIES AND TO BECOME A SCIENTIST. YOU TAUGHT ME HOW TO TEACH MYSELF ANYTHING. I WISH YOU COULD HAVE SEEN THIS WORK COMPLETED. I WISH YOUR ATOMS A HAPPY NEW CYCLE, INTERMINGLING WITH THE OLIVE TREES OF YOUR BELOVED GARDEN.

Acknowledgments

THERE ARE TOO MANY PEOPLE TO THANK, REALLY! But I will try in this short text to do so.

First, I want to thank all the members of the Extavour laboratory. The days when going to the lab was hard, knowing that such a wonderful community of people were there always cheered me up. But more importantly, everyone taught me so much, about so many things. In particular, Seth Donoughue, thank you for always having crazy ideas and setting such a great example for me to follow. Taro Nakamura, for sharing your knowledge and always being happy to help me troubleshoot any experiment. Tarun Kumar, for being an amazing partner in science, I will always remember our collaboration fondly.

I also want to thank the members of the Harvard community whom I met and shared many adventures with. I was blessed with Ph.D. classmates who were so open-minded and kind. You helped me with science and life, and I am so grateful you were such a beautifully weird group. Thank you Finn, Sean, Georgia, Jamilla, Felix, Alyson, Katie, Yiqun, Jun Han, Linda, Nico, Matt & Sam. And I want to thank the wonderful musicians from around the world with whom I shared amazing musical moments. Harvard World Music Ensemble friends, thank you for the wonderful musical adventures we lived together. I was blessed by the chance to learn from all of you.

To all the staff that make the research endeavour at Harvard possible. Thank you Jack Conlin, Patty Gonzalez and Mike Lawrence for all your dedicated work to help the MCO community. Claire Reardon and Douglas Richardson for all the time you spent helping me with convoluted scientific apparatus. The building operation and custodial service staff for being so kind and helpful.

This adventure was also supported outside of Harvard by the members of the community I lived in during my Ph.D., the Ridgemonites. Thank you for the adventures, food, music, movies, trips, debates, discussions, and so much more. Thank you Marc, Juan, and Deborah for the time we spent together. Thank you Marc for all the knowledge of networks, physics, math, and music you gave me. Thank you Deborah for inviting me to your Philosophy groups at MIT. Thank you Juan for teaching me the ins and outs of business and life hacks.

And from as early as I can remember, thank you my family who always supporting me and accepted me as I was. My mom, Marie-Claude Larrondo who has lived the life of an academic without becoming one.

Your curiosity and drive to never stop learning still lives in me. My dad, Vincent Blondel, who's perseverance and love for sciences is one of the main reasons I was able to do a Ph.D. Thank you both for your unwavering love and support. My sister, Lisa Blondel, who's creativity is infectious, and probably the only person I can communicate hours-long discussions without a single word being spoken. My partner, Brenda Marin Rodriguez, for the most wonderful love and support one could have ever hoped for.

I want to thank the member of my Ph.D. committee. The directions and help I received from Allon Klein, Craig Hunter and Sharad Ramanathan were instrumental in the completion of this degree. Thank you for the time you took to listen, and point me in directions I would not have conceived on my own.

And finally, I want to thank my advisor, Cassandra Extavour. When I joined your lab, I knew that it was going to be hard. You are the most rigorous person I know, and I was maybe one of the least rigorous students you had, but your patience to teach me, point my mistakes, and help me grow into a scientist is what molded me into who I am today. Thank you for always having direct conversations, never sugarcoating my mistakes. Thank you for becoming a mentor, in life and in science.

*Se tenir sur les épaules des géants et voir plus loin. Voir dans l'invisible, à travers l'espace et à travers le temps. Plonger notre regard dans le passé et découvrir que notre passé est immense. Pouvoir remonter le temps à contre courant. Pouvoir distinguer à travers le long écoulement des âges, des éclats de passé qui soudain, resurgissent de l'oubli. Des éclats de mondes disparus. Et partir à la recherche des lointaines métamorphoses qui ont donné naissance au monde d'aujourd'hui. **

Jean-Claude Ameisen

O

Introduction

*

*Stand on the shoulders of giants and see further. To see in the invisible, through space and through time. To plunge our gaze into the past and discover that our past is immense. To be able to go back in time, against the current. To be able to distinguish, through the long flow of ages, bursts of the past that suddenly resurface from oblivion. Shards of vanished worlds. And go in search of the distant metamorphoses that gave birth to the world of today.

The cells responsible for the transmission of the genetic material to the next generation are called germ cells. Germ cells are a specialized cell lineage that is capable through division and differentiation of generating all cell types of the organism^{1,2}. Germ cells are speculated to have stemmed from a unique ancestral stem-cell population^{2,3} (reviewed by Ewen-Campen et al.⁴), however the way that they are specified during embryogenesis differs across organisms. Animal germline specification can happen in one of two ways, often termed "inheritance" (through the use of localized determinants) and "induction"⁵. The inheritance mechanism starts with the localized deposition of maternal RNA and protein germ cell determinants in the oocyte. Then, through asymmetric inheritance during the first embryonic divisions, a subset of the embryonic cells inherit the deposited material. Those cells then acquire primordial germ cell fate and give rise to the germline. In the induction mechanism, the specification of germ cells happens (often in the embryo) through signals sent by somatic cells to the future germ cells. A small population of cells receive and react to this signaling and transforms the primordial germ cell population (reviewed by Ewen-Campen et al.⁴). Whether all embryonic cells or a competent population of cells can react to those signals depends on the organism (reviewed by Extavour and Akam⁵, Seervai and Wessel⁶). But, the bi-modality of germ cell specification has been disputed, and it has been proposed that this phenomenon might instead be a more continuous process that re-uses the same components (reviewed in Ewen-Campen et al.⁴, Seervai and Wessel⁶). Indeed, many of the genes present in the germ cell localized determinants have also been found to be under regulation by inductive mechanisms (reviewed by Seervai and Wessel⁶). Moreover, multiple organisms display specification strategies that utilize both an inheritance and an inductive mechanisms. For example, in the wasp *Pimpla turionellae*, when the Oosomes (the necessary germ cell localized determinants⁷) are removed during early embryogenesis, the adults will compensate for this loss and specify a functional germ cell population⁸. In the sea urchin, an organism thought to exhibit an induction strategy⁹, experiments separating early blastomeres found that only the vegetal blastomere retained the capacity to form germ cells, hinting at the presence of a localized determinant¹⁰. Finally, the numerous transitions from one specification strategy to the other⁵ could be indicative of a single continuous process where evolution acts on a dial that changes from fully inherited to fully inductive with all other possible states in between (reviewed by Seervai and Wessel⁶).

The mechanism of germline specification has been extensively studied in *Drosophila melanogaster*. This organism uses an inheritance mechanism, where the asymmetric deposition of germline determinants (called germ plasm) leads to the formation of the primordial germ cells^{11,12,13,14,15}. However, the inheritance mechanism does not appear to be the ancestral mechanism for germline specification in insects. In basally branching insects, no germ plasm has been reported, and germ cells appear to be

specified through an inductive mechanism¹⁶. The hypothesized shift from induction to inheritance required the acquisition of a new function, the asymmetric deposition of germ plasm. In *D. melanogaster*, a gene of initially unknown origin called *oskar* was found to be necessary and sufficient for the localization of germ plasm and specification of germ cell fate^{11,14,15}.

0.1 The origin and evolution of *oskar*

In *D. melanogaster*, cells at the posterior end of the embryos that inherit a cellular component called germ plasm will become the primordial germ cells and are called pole cells¹⁷. As shown by early transplantation experiments, this germ plasm is necessary and sufficient to induce the formation of germ cells by cells that inherit it¹⁷. This mechanism requires the transport of germ plasm components to the posterior pole of the oocyte^{18,19,20}. A key component required for the assembly of this germ plasm is the gene *oskar*^{11,14,15}. The translation of *oskar* mRNA is inhibited until it has reached the posterior pole of the embryo, upon which it will be translated into two isoforms, Long and a Short Oskar²¹. The isoforms differ by the addition of 138 amino acids at the N-terminal domain of Long Oskar²¹. Interestingly, this seemingly small addition led to a very different function of the Short and Long Oskar isoforms. While Short Oskar is necessary for the formation of pole cells (the primordial germ cells of *D. melanogaster*)²¹, it does not participate in the anchoring of the germ plasm at the posterior pole²². Long Oskar, however, is essential for the localization of the germ plasm at the posterior pole, but cannot induce the formation of germ plasm²². The mechanism by which *oskar* induces the formation of germ plasm is still unknown. Interestingly, other organisms specifying their germ cell formation through an inheritance mechanism seem to possess a similar nucleator as *oskar* (reviewed in Kulkarni and Extavour²³). In *Danio rerio* (Zebrafish), the gene *bucky ball* is necessary and sufficient to organize the Balbiani body²⁴. In *Caenorhabditis elegans*, two genes, *pgl-1* and *pgl-3*, were found to be the nucleator of the P-granules, a key component of germ cell specification^{25,26}. To better understand how such mechanisms of specification could arise, in the first and second chapters I studied the origin and evolution of the gene *oskar*.

0.2 The origin of new genes

In the first chapter of this dissertation, I focus on uncovering the evolutionary origin of *oskar*. As mentioned above, *oskar* is a key determinant in the specification of germ cells in *D. melanogaster*^{11,14,15}. Many holometabolous insects specify their germ cells through the inheritance of maternally deposited germ plasm (discussed in Lynch et al.²⁷). However, holometabolous species such as *Apis mellifera*²⁸ or

*Bombyx mori*²⁹ do not display a similar mechanism. In both species *oskar* is seemingly absent from their genome²⁷. Lynch et al.²⁷ proposed that the absence of *oskar* in holometabolous species correlates with the absence of an inheritance-mediated germline specification mechanism. Despite its central role, no homologs of *oskar* have been found outside of insects^{16,27}. Therefore, I wanted to know how a gene that does not seem to have homologs predating the evolution of insects became so central to their reproduction. Was *oskar* a new gene? And if it was a new gene, what evolutionary changes led to its functions?

Genesis mechanisms and importance of new genes

The first proposed mechanism of new gene evolution was duplication, predicted to be followed by relaxed selection and mutation of the duplicated copy (reviewed by Kaessmann³⁰, Innan and Kondrashov³¹). Jacob³² proposed that evolution does not have a plan, it does not engineer but tinkers with already existing parts. One such tinkering mechanism proposed for the formation of new genes is called rearrangement (reviewed by³⁰). Rearrangement is the reordering of parts of genes in a new order that gives rise to new functions (reviewed by Kaessmann³⁰). However, the duplication and rearrangement mechanisms fail to explain the existence of so-called new genes (also called orphan genes or novel genes), genes with no homologs found outside a given lineage (reviewed by Tautz and Domazet-Lošo³³). Based on the reported lack of homologs inside and outside of insects³, *oskar* can be described as a new gene. One of the main mechanisms of the formation of new genes is by *de novo* transcription and translation of previously non-coding regions of the genome (reviewed by Tautz and Domazet-Lošo³³). Under this mechanism, a region of the genome would acquire a transcription starting site through mutation. The RNA molecule produced could function as a non-coding RNA until an open reading frame (ORF) emerged within its sequence, again by random mutation. While the multiple rare events involved in this mechanism seems unlikely, a number of cases of *de novo* evolution have been documented^{34,35,36,37,38}. Another mechanism by which organisms can acquire new genes is through the transfer of DNA from another organism and subsequent integration of this DNA into their genome. This process is called horizontal gene transfer and it has been speculated that this process can drive the acquisition of new functions in eukaryotes^{39,40,41,42,43}. In the *Drosophila* subgroup, a study estimated the rate of origination of new genes to be between 5 and 11 new genes per million years of evolution⁴⁴. A majority (55%) of those genes became fixed in the drosophilids⁴⁴. A later study estimated that in the *Drosophilid* lineage, the rate could be as high as 17 new genes for every million years⁴⁵. Interestingly, this rate is similar to the rate of gene loss, leading to an apparent stability in total gene number, with older genes being replaced by new genes⁴⁵. Not only do new genes appear at a higher rate than was previously

envisioned^{46,47}, they can also play key roles in the acquisition of new functions. The protein product of new genes might be under relaxed selection allowing it to find new interactors and participate in pre-existing biological functions or generate novel functions altogether (reviewed in Tautz and Domazet-Lošo³³). Essential genes are the subset of genes without which the ability of an organism to grow and reproduce is compromised (reviewed by Lewin et al.⁴⁸). Given their importance, it was hypothesized that they must be conserved and ancient (reviewed by Lewin et al.⁴⁸). However, in *D. melanogaster* 30% (59 of 195) new genes (less than 35 million years old) were found to have a lethal phenotype under a RNAi knockdown⁴⁹. This proportion is similar to the 35% (86 of 245) found for a random sample of old genes lethal under a knockdown⁴⁹. Given the similar proportion, the authors hypothesized that essentiality was not a hallmark of old genes⁴⁹. The age of new genes is also correlated with their centrality in the human and mouse gene-gene interaction (GGI) networks⁵⁰. Younger new genes tend to have a low centrality and be positioned at the periphery of the GGIs, while older new genes are found towards the center of the network with a high centrality and hub topologies⁵⁰. Younger new genes also have a higher rate of partner acquisition than older new genes⁵⁰. Therefore it has been hypothesized that new genes act as a driver of topological change in GGI networks⁵⁰. Taken together with the reports that genes with a higher centrality have a higher chance to be essential genes⁵¹, I suggest that essentiality is a dynamic process whereby new genes emerge, then integrate within existing networks until a proportion become central hubs, and therefore essential. Finally, it has been hypothesized that new genes tended to be expressed more frequently in specific tissues such as the brain and testis⁵². For example, in *D. melanogaster*, almost half (48.8%) of the genes that originated since the divergence with *D. pseudoobscura* were found to be expressed in the brain⁵³. A subset of those genes was found to have stereotypic expression patterns in specific neuronal populations, which was hypothesized as a marker of functionalization⁵³. However, an alternate hypothesis could be that the diversity of neuronal cell types increase the propensity for diverse gene expression, or that neuronal cells are more permissive to new gene expression.

Oskar domains, LOTUS and OSK display differential sequence identities

If we disregard the dipteran-specific Long Oskar isoform, the structure of Oskar is composed of two conserved domains interspaced by an unconserved region^{54,55}. The first domain is a winged-helix domain called LOTUS and is found across eukaryotes, including within the Tudor domain-containing protein family⁵⁶. The second domain called the OSK domain, however, appears to have no homolog within eukaryotes and shows a high sequence and structural similarity to bacterial GDSL lipases^{27,54,55}. According to the descriptions of new gene formation mechanisms described above, *oskar* did not seem to

fit any of the classic duplication, rearrangement, or *de-novo* mechanisms. However, the similarity between the OSK domain and bacterial sequences led Lynch et al.²⁷ to hypothesize that it might be the product of a horizontal gene transfer. In insects, multiple reports have documented endosymbiosis with bacteria (reviewed in⁵⁷). Bacteria of the *Wolbachia* family are an essential and required symbiont for the reproduction of multiple wasp species⁵⁸. In multiple insect species, pieces, or in some cases the entirety, of the *Wolbachia* genome were found integrated into the nuclear genomes, which was interpreted as resulting from a bacterial endosymbiosis⁵⁹. *Acyrtosiphon pisum*, an Aphid species, is dependent on the capacity of their endosymbionts to produce essential amino acids that they cannot receive from the sap they feed on (reviewed by Oliver et al.⁶⁰). If a horizontal gene transfer event was sufficiently recent, it is often possible to detect its signature from a shift in GC content in a region of the genome, or through careful calculation of codon frequencies in the case of a protein-coding gene⁶¹. But to provide evidence that a new gene is the product of a horizontal gene transfer, the most thorough methodology is careful phylogenetic analysis⁶¹. Therefore, in the first chapter, I set out to understand the evolutionary origin of the gene *oskar*, through careful phylogenetic analysis, focusing on testing the hypothesis for a partial gene transfer (or horizontal domain transfer).

0.3 The evolution of *oskar*

As described above, the absence of *oskar* in an insect genome correlates with the absence of a localized determinant of germ cell specification²⁷, but the reverse is not true. In *Gryllus bimaculatus* (cricket), the gene *oskar* is present in the genome and expressed in embryonic neuroblasts and adults' brains⁶². However, *G. bimaculatus* specifies its germline through an induction mechanism that does not require *oskar*⁶³. This suggests that during the evolution of insects, *oskar*, as other new genes did^{50,53}, first neofunctionalized within the embryonic nervous system. In holometabolous insects, however, the gene *oskar* became essential for germ cell specification²⁷. It was therefore hypothesized that *oskar* was co-opted and acquired a new function, namely to nucleate and localize the germ plasm at the posterior pole of some, but not all, holometabolous insects²⁷ to specify germ cells⁶². The order of events of *oskar* functionlization is based upon the absence of localized gem cell determinant in hemimetabolous insects, however, without more studies of the mechanism of germ line specification in hemimetabolous insects, I cannot rule out that *oskar* acquired a function in germ cell specification prior to the split with holometabola. While our understanding of the function of *oskar* in insects remains poor, in *D. melanogaster* a significant amount of experimental information has been documented about its biochemistry and structure. The Oskar protein interacts with multiple proteins involved in the regulation

and specification of the germline. Oskar associates with Vasa^{56,64}, a DEAD-box helicase conserved throughout many animals as a marker of germ cells in early embryogenesis (reviewed in Ewen-Campen et al.⁴, Extavour and Akam⁵, Noce et al.⁶⁵, Raz⁶⁶). Oskar was found to interact with Staufien in yeast two hybrid and *in vitro* pull down experiments⁶⁴ and Valois in *in vitro* pull down experiments⁶⁷, two proteins involved in the formation of germ plasm^{12,68}. Of note, Oskar interacts with its own mRNA, through a mechanism involving the binding of its 3'UTR⁵⁵. Moreover, *oskar* alleles with point mutations in the OSK domain prevent the maintenance of the *oskar* mRNA at the posterior pole, despite its correct production and localization during the formation of the oocyte⁶⁸. Those alleles break the interaction between *oskar* mRNA and the OSK domain⁵⁵. *nanos* is a conserved gene involved in the specification and localization of the germline^{4,5}. The 3'UTR of *nanos* mRNA also binds to the OSK domain, and alleles that disrupt the 3'UTR binding and localisation of *osk* mRNA also prevent the binding and localisation of *nanos* mRNA⁵⁵. As well as the specific binding of *nanos* and *oskar* mRNA, OSK also appears to be a general RNA binding domain⁵⁴. Finally, despite its similarity to SGNH or GDSL lipases^{54,55}, the conserved triad of catalytic amino acids present in those classes of enzyme is absent from the OSK domain^{54,55}. To my knowledge, apart for the OSK domain no RNA binding has been reported for SGNH or GDSL lipases.

In *D. melanogaster*, when expressed in isolation, the LOTUS domain of Oskar can dimerize⁵⁴. It is believed to homodimerize through an electrostatic and hydrophilic interface composed of the β 2 strand of its β -sheet and α -helix α 4⁵⁴. However, when artificially expressed, LOTUS domains of other insect species (two other dipteran species, five hymenopteran species and one orthopteran species) displayed either a monomeric (6 out of 9 tested species) or a dimeric state (3 out of 9 tested species), as assayed by static light scattering⁵⁴. Therefore, it was hypothesized that the dimerization property of the LOTUS domain was not conserved within all insects⁵⁴. Moreover, the LOTUS domain interacts directly with Vasa through an interface composed of the α -helices α 2 and α 5 of the LOTUS domain⁵⁶. This interaction increases the helicase activity of Vasa⁵⁶. The LOTUS domain has also previously been predicted to be an RNA binding domain^{69,70}, and its structure aligns closely with that of MecI, a dsDNA binding domain⁵⁵. While the LOTUS domain does not bind the 3'UTR of *oskar* mRNA⁵⁵, no experiments outside of *D. melanogaster* have been performed to support or contradict the RNA or dsDNA binding prediction.

Expanding on the conclusions of the first chapter, I wanted to know what evolutionary changes in the sequence of *oskar* could have led to the acquisition of its germ plasm nucleation capacity in holometabola. In the second chapter, we therefore analyzed the evolutionary sequence changes of 379 sequences of *oskar*, sampled from a majority of insect orders, to try to uncover the changes that happened between hemimetabolous insects and holometabolous insects. We also explored the conservation of specific residues in view of the known functions of *oskar* and proposed hypotheses for

new important residues.

0.4 The regulation of *D. melanogaster* ovariole formation and egg-laying

The fate of germ cells is intricately linked to the development of reproductive organs in animals. In *D. melanogaster*, primordial germ cells are internalized inside the embryo and migrate towards the location of the formation of the embryonic ovaries and testes^{71,72}. In insects, ovaries are organized into repeating structures called ovarioles^{73,74}. This structure contains the oocyte production machinery (reviewed by Wheeler⁷³) and is of a tubular shape where the oocyte starts its formation at one end and exits it as a mature egg at the other end^{73,74}. At the anterior end of the ovariole is a structure containing a stack of disk shaped cells called the terminal filament. Posterior to the terminal filament is a structure called the germarium which contains the germ cells^{73,74}. The number of ovarioles in insects can vary greatly even within the same insect order⁷⁵. For example, in Coleoptera, *Meloe proscarabaeus* has up to a thousand ovarioles⁷⁶, whereas members of Scarabaeinae have only one⁷⁷. During the development of the ovaries in *D. melanogaster*, terminal filament cells organize themselves into a stack which then becomes the terminal filament (reviewed by Wheeler⁷³). Because each ovariole develops from a stack of terminal filament cells, the number of terminal filaments formed in the developing ovary can be used to predict the number of ovarioles formed in an adult ovary^{73,74}. It has been hypothesized that the number of ovarioles informs the reproductive capacity of insects by modulating their capacity to lay a set number of eggs, and has been found to be highly variable and to display phenotypic plasticity (reviewed in Hodin⁷⁸). While the number of ovarioles in *D. melanogaster* can vary when exposed to different environmental conditions, including temperature⁷⁹ or altitude and climate^{80,81}, in a constant environment it is highly stereotypical⁸². This stereotypicality implies a strong genetic regulation of the number of terminal filaments. For example, one QTL analysis of genomic variation in the Drosophila Genetic Reference Panel (DGRP, a collection of 200 inbred lines) lines of *D. melanogaster* revealed loci strongly correlated with variation in ovariole number⁸³. The control of the specification of terminal filament cells in the ovary plays a key role in determining ovariole number⁸⁴. The *hippo* signaling pathway, involved in the regulation of organ size and cell proliferation (reviewed by Seb e-Pedr os et al.⁸⁵, Dong et al.⁸⁶), was previously demonstrated to control the proliferation of terminal filament cells⁸⁷. Under a repression of *hippo* signaling in the developing ovary, the number of terminal filament cells increases, along with the number of terminal filaments and ovarioles formed in the adult ovary⁸⁴. Further experiments showed

that an over-expression of *yorkie*, the main effector of *hippo* signaling, induced a decrease in the number of terminal filament cells, terminal filaments, and ovarioles⁸⁷. In the third chapter of this thesis, we expanded on those previous results and explored the role of all signaling pathways in the regulation of ovariole number and of reproductive capacity measured as the number of eggs laid in a given time frame. Extavour lab postdoc Tarun Kumar performed a knockdown screen of all known members of signaling pathways in the developing ovary and recorded each phenotypic output. To analyze this dataset, we decided to use tools developed in the field of systems biology and network science. We set out to answer questions such as: What is the contribution of each signaling pathway to the number of ovarioles and egg laying capacity? Are those signaling pathways working in modules? Are there signaling pathway-independent regulatory modules that control ovariole number and egg laying? Can we find epistatic relations between other signaling pathways and *hippo* signaling?

Genes, proteins, and other molecules are connected to each other by their regulatory capacities. An abstraction of the mechanisms involved in the regulation of cells and organisms is found in network science^{88,89}. By modeling an entity such as a gene, or its protein product, as a node, and their interactions as edges connecting both in a directional or non-directional manner, it becomes possible to understand higher order regulatory structures^{90,91}. One of the concepts to describe and understand how hundreds of genes might interact together is that of modularity, or gene regulatory modules^{90,92,93}. A module has been defined in multiple ways⁹⁴ but in the third chapter, I focus on the definition given by researchers working with the concept of disease modules^{92,95,96}: by this definition, a module is a group of genes involved in controlling a phenotype and showing a statistical enrichment in topological features. The topology of graphs can be described by multiple features, including but not limited to the number of edges connecting a subset of nodes, also called density; the centrality of those nodes with regard to the rest of the graph, or a subset of the graph; the size of the largest connected component; and the average shortest path length between all nodes in the subset (reviewed by Barabási⁹⁷). Using those features and phenotypic data allowed, for example, the discovery of a set of protein interactions regulating the intensity of a placebo effect⁹⁶. Topological approaches to studying the regulation of cell differentiation also allowed for the discovery of regulatory motifs in the early embryogenesis of *D. melanogaster*⁹⁸. However, while modularity in itself is important for functional separation of parts of the network, other topological features are important for the resilience of those functions to evolutionary pressures⁹⁹.

Using those approaches, we discovered distinct sub-networks of signaling pathway genes regulating the number of ovariole as well as the egg laying capacity of *D. melanogaster*. Those sub-network shared a common subnetwork of genes which influenced heavily both phenotypes. Our analysis also revealed that the control of both phenotypes was under the regulation of all signaling pathway, at least to some extent.

Finally, our analysis successfully predicted the involvement of previously uncharacterised genes with higher accuracy than the original candidate screen.

0.5 The unique features of crustacean early embryonic development and germ layer specification

In the fourth and fifth chapters, I studied the development of another organism, *Parhyale hawaiiensis*. *P. hawaiiensis* is a crustacean amphipod. Amphipods are members of a very diverse group called the Malacostraca (reviewed by Thiel and Wellborn¹⁰⁰). The Malacostraca includes 16 crustacean orders and over 40000 species (reviewed by Thiel and Wellborn¹⁰⁰). Animals we generally refer to as crustaceans are found within the Malacostraca, such as Decapoda (lobsters, crabs, and shrimp), Peracarida (isopods, amphipods), and Stomatopoda (mantis shrimp) (reviewed by Thiel and Wellborn¹⁰⁰). From a developmental perspective, animals in the Malacostraca display a number of unique features reviewed in Wolff and Gerberding¹⁰¹, Scholtz and Wolff¹⁰² such as a unique nauplius stage, organisms with direct development, multiple form of invariant lineages and variations between holoblastic and superficial cleavages.

In the malacostracan organisms showing invariant early cell lineages, the specification of the germ layer is restricted by lineages. However, the division patterns between Amphipoda, Euphausiacea, and Decapoda are very different, implying different germ layer specification strategies (reviewed by Scholtz and Wolff¹⁰²). It is therefore likely that molecular mechanisms linked to the division pattern instruct the specification of the germ layers. To my knowledge, the only malacostracan organism where molecular mechanisms of the specification of germ layers has been studied is *P. hawaiiensis*, where a study showed asymmetric inheritance of mRNA by the different precursor of each germ layer¹⁰³. Despite the lack of molecular studies, the elucidation of the invariant lineages in crustaceans has a long history and started as early as 1879 with the description of the cleavage pattern of *Moina rectirostris*¹⁰⁴ to modern studies in *Orchestia cavimana*¹⁰⁵ and *P. hawaiiensis*¹⁰⁶. Within the Malacostraca, embryos of Decapoda, Dendrobranchiata, Euphausiacea, and Amphipoda display an invariant early cell division (discussed in Gerberding et al.¹⁰⁶). While the number of early cell divisions that occurs before the specification of each germ layer varies between malacostracan orders (from seven divisions in *Penaeus* and *Sicyonia* (Decapoda) to three in *Parhyale* and *Orchestia* (Amphipoda) (discussed in Gerberding et al.¹⁰⁶) the specification of germ layers happens at an earlier cell division cycle compared to other arthropods such as *D. melanogaster* where this process happens at the 14th division (reviewed by Gilbert¹⁰⁷). In

amphipods, the first three cell divisions lead to the formation of an asymmetric embryo composed of eight blastomeres, four large macromeres on the ventral side and four small micromeres on the dorsal side^{105,106}. Three of the macromeres form the ectoderm, one macromere and two micromeres compose the mesoderm, one micromere the endoderm and the eighth micromere gives rise to the germline^{105,106}. In the fourth and fifth chapters, I focused on trying to uncover the molecular mechanisms underlying the specification of the germ layers in *P. hawaiiensis* using single-cell RNA sequencing and light sheet microscopy.

0.6 The molecular basis of *P. hawaiiensis* germ layer formation

Recently, multiple laboratories have developed new methodologies that allow for the sequencing of the RNA content of a single cell^{108,109,110,111,112}. To increase the multiplexing capacity of single-cell RNA sequencing drastically, Drop-Seq and inDrop^{109,110} use a strategy that encapsulates cells in a micro reaction chamber using microfluidic devices. While the advantages of single-cell RNA sequencing allows the obtention of a very high-resolution cartography of the gene expression in a developing embryo, there exist many challenges to successfully sequence the cellular contents which are reviewed in Denisenko et al.¹¹³.

In the fourth chapter, I describe my attempts to use single-cell RNA sequencing to understand *P. hawaiiensis* early embryogenesis. I aimed to determine the transcription profiles of the well defined cell lineages coming from the eight blastomeres. My ultimate goal was to combine this single-cell RNA sequencing data with single-cell resolution imaging of early embryos that I aimed to collect as described in the fifth chapter.

0.7 *P. hawaiiensis* early embryogenesis

While experiments pertaining to the molecular mechanisms of germ layer identity in Malacostracan are sparse (see above for more detail), information on the morphogenesis of malacostracan embryos has been documented for many species (reviewed by Wolff and Gerberding¹⁰¹). Malacostracan embryos can follow a variant or invariant division pattern (reviewed in Gerberding and Patel¹¹⁴), but recent descriptions of early cleavages and cell lineages have focused primarily on invariant embryos^{106,115,116,117,118,119}. The study of invariant cleavages and gastrulation has been documented for

species such as the krill species *Meganyctiphanes norvegica* (Euphausiacea)¹¹⁵, the shrimp *Penaeus vannamei* (Decapoda)¹¹⁶, *P. monodon*¹¹⁸, *P. japonicus*¹²⁰ and *Sicyonia ingentis*¹¹⁷, and the amphipods *Orchestia cavimana* (Amphipoda)¹²¹, and *P. hawaiiensis* (Amphipoda)¹⁰⁶. While all three groups display invariant holoblastic early divisions, their cleavages and morphogenetic events are very different (reviewed by¹⁰¹). In Amphipoda, the first three cleavages result in an asymmetry between the macromeres on the ventral side and the micromeres on the dorsal side^{106,119}. In Decapoda, the first three cleavages produce cells that are slightly unequal in size, but in comparison to the asymmetry in Amphipoda, the eight blastomeres are equally distributed and of relatively similar sizes^{116,118}. In Euphausiacea, the first three cleavages result in eight blastomeres of equal size¹¹⁶. Despite their very different cleavage dynamics, in all three groups, the cells that will form the ectodermal lineages undergo, subsequently to their specification, a more rapid division cycle than the other cells^{116,118,119,122}.

The gastrulation of crustaceans exhibits a very high level of variation (discussed by Gerberding and Patel¹¹⁴). In Euphausiacea^{115,123,124} and Decapoda^{116,117,118,125}, gastrulation happens via cellular ingression at the vegetal pole (reviewed in Gerberding and Patel¹¹⁴). At the 7th division cycle (128 cell stage) the future mesoderm and endoderm cells are organized in a ring shape around the vegetal pole, marked by the intra-cellular body (ICB), which is hypothesized to be a germ plasm equivalent in Decapoda¹²⁰. Between the 6th and 7th division cycle the cellular divisions of eight and then 16 cells form a radially oriented rosette towards the vegetal pole^{115,116,117,118}. This rosette is hypothesized to be the driver of the ingression towards the interior of the blastula (discussed in Gerberding and Patel¹¹⁴).

In Amphipoda, gastrulation happens at the anterior pole of the embryo instead of the posterior pole (discussed in Gerberding and Patel¹¹⁴). By the eight cell stage, the fate of the micromeres and macromeres is already established^{106,119}. In *Orchestia cavimana* and *P. hawaiiensis*, gastrulation was reported to be very similar, though with one key difference. In both species, the second cleavage ancestor of the visceral mesoderm and germline (*A/a* in *Orchestia cavimana* and *Mav/g* in *P. hawaiiensis*) is covered by the ectodermal cells through an active epibolic process^{126,127}. In *P. hawaiiensis*, the descendants of *ml* and *mr* (the micromeres fated to give rise to the somatic mesoderm, homologous to *b* and *d* in *Orchestia cavimana*) migrate away from the anterior pole and later ingress inside the embryo. In contrast, *Orchestia cavimana* *ba* and *da* cells (the anterior descendants of *b* and *d*, homologs of *mla/mra* in *P. hawaiiensis*) ingress inward at the same site as the descendants of *a* and *A*. In *Orchestia cavimana*, the descendants of *a* (germline precursor) ingress before all other cells¹²⁷. It has been hypothesized that they are the main driver of gastrulation¹²⁷. However, in *P. hawaiiensis*, *g* and *Mav* ingress independently of each other¹²⁶. Given that *Mav* and *g* are independently ingressing, the authors of this study questioned the validity of *a* as the main driver in *Orchestia cavimana* and proposed that a similar mechanism could be at play

(discussed in Chaw and Patel¹²⁶). However, I believe that this might also be a difference between the two species.

In the fifth chapter, I wished to study the dynamic of early embryogenesis in *P. hawaiiensis*. More specifically, what are the dynamics of the cellular lineage territories? Given the invariance of early cleavages¹⁰⁶ and stereotypic germ band stage¹²⁸, what is the variance in cellular movement and positions between both stages? What cellular movement and cellular rearrangement lead to the formation of rows at the germ band stage? To tackle those questions, I used light-sheet microscopy^{129,130} to image live developing *P. hawaiiensis* embryos from the eight (and 16) cell stage to the germ band extension stage. To track the cells and assess cellular rearrangements, the nuclei and cellular membranes were tagged with fluorescent markers. Upon ablation of a blastomere from the ectodermal or mesodermal lineages, embryos display an intra germ layer compensation mechanism¹³¹. Upon ablation of the right (Er) or left (El) ectodermal blastoderm, cells from the posterior and remaining lateral ectodermal lineages replace the missing cells¹³¹, breaking two stereotypic barriers: the anterior-posterior boundary and the midline. By ablating Er or El, and subsequently recording the development of the embryo, I hoped to generate hypotheses towards the mechanisms establishing both boundaries. Finally I also hoped to be able to combine these imaging data with the single-cell RNA sequencing dataset that was the goal of the fourth chapter, by using the light-sheet dataset as a reference atlas for the geometrical mapping of gene expression in the developing embryo.

0.8 The interactive visualization of microscopy images

In the last chapter I describe the creation of a Virtual Reality (VR) tool to observe 4D microscopy datasets and track nuclei in 3D. When confronted with the visualization of three-dimensional datasets, computer rendering software will reproject the 3D images onto a two-dimensional screen, therefore losing one of the original dimensions. Using the concept of stereography, augmented reality CAVE (Cave Automatic Virtual Environment) systems were invented to circumvent that issue and observe the specimen in 3D¹³². However, those are very costly, require specialized hardware, and occupy a large amount of space¹³². With the invention of consumer virtual reality headsets, we now can display those images in 3D, for a relatively low price¹³³. Moreover, with "room-scale" virtual reality and controllers projected into the virtual environment, it is not only possible to observe, but also to manipulate the samples. In the last chapter of this thesis, I explore the new possibilities for the observation of 4D biological datasets offered by Virtual Reality. To visualize the datasets generated in the fifth chapter, I built the foundations of a VR software that allows the user to import any three-channel volumetric movie and perform image

adjustments, slicing, and playing/pausing. Using this tool allows for new insights and an intuitive understanding of the embryos observed. However, while the observation of 4D datasets was completed, the creation of tracking features was not.

One of the key aspects of the fifth chapter was the tracking of nuclei in the developing embryo of *P. hawaiiensis*. To this end, I used the light-sheet dataset FIJI tracking plugin Mamut¹³⁴. However, predicting the movement of nuclei in 3D on a 2D screen is complex, and more often than not, I needed to search for a lost nucleus that went out of the rendered plane. By projecting the embryo in three dimensions in a Virtual Reality environment, the observation of objects becomes continuous, no slices are taken as the object in its integrity is itself in front of the user. A 3D tracking software in VR that would then allow the user to continuously track nuclei in 3D space by using the Virtual Reality controllers.

References

- [1] J. Srouji and C. G. M. Extavour. Redefining stem cells and assembling germ plasm:. In R. Desalle and B. Schierwater, editors, *Key transitions in the evolution of the germ line*, page 360. CRC Press, 2011.
- [2] C. G. M. Extavour. Evolution of the bilaterian germ line: lineage origin and modulation of specification mechanisms. *Integr. Comp. Biol.*, 47(5):770–785, November 2007.
- [3] C. G. M. Extavour. Gray anatomy: phylogenetic patterns of somatic gonad structures and reproductive strategies across the Bilateria. *Integr. Comp. Biol.*, 47(3):420–426, September 2007.
- [4] B. Ewen-Campen, E. E. Schwager, and C. G. M. Extavour. The molecular machinery of germ line specification. *Mol. Reprod. Dev.*, 77(1):3–18, 2010.
- [5] C. G. Extavour and M. Akam. Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development*, 130(24):5869–5884, December 2003.
- [6] R. N. Seervai and G. M. Wessel. Lessons for inductive germline determination. *Mol Reprod Dev*, 80(8):590–609, Aug 2013.
- [7] J. F. Bronskill. EMBRYOLOGY OF PIMPLA TURIONELLAE (L.) (HYMENOPTERA: ICHNEUMONIDAE). *Canadian Journal of Zoology*, 37(5):655–688, 1959. doi: 10.1139/z59-068. URL <https://doi.org/10.1139/z59-068>.
- [8] M. Achtelig and G. Krause. [Experiments on the uncleared egg of *Pimpla turionellae* L. (Hymenoptera) for the functional analysis of the oosome region]. *Wilhelm Roux Arch Entwickl Mech Org*, 167(2):164–182, Jun 1971.
- [9] S. Hörstadius et al. Experimental embryology of echinoderms. 1973.
- [10] E. Voronina, M. Lopez, C. E. Juliano, E. Gustafson, J. L. Song, C. Extavour, S. George, P. Oliveri, D. McClay, and G. Wessel. Vasa protein expression is restricted to the small micromeres of the sea urchin, but is inducible in other lineages early in development. *Dev Biol*, 314(2):276–286, Feb 2008.

- [11] A. Ephrussi and R. Lehmann. Induction of germ cell formation by *oskar*. *Nature*, 358(6385):387–392, July 1992.
- [12] A. P. Mahowald. Assembly of the *Drosophila* germ plasm. *Int. Rev. Cytol.*, 203:187–213, 2001.
- [13] A. C. Santos and R. Lehmann. Germ cell specification and migration in *Drosophila* and beyond. *Curr. Biol.*, 14(14):R578–89, July 2004.
- [14] J. L. Smith, J. E. Wilson, and P. M. Macdonald. Overexpression of *oskar* directs ectopic activation of *nanos* and presumptive pole cell formation in *Drosophila* embryos. *Cell*, 70(5):849–859, September 1992.
- [15] R. Lehmann and C. Nüsslein-Volhard. Abdominal segmentation, pole cell formation, and embryonic polarity require the localized activity of *oskar*, a maternal gene in *Drosophila*. *Cell*, 47(1):141–152, October 1986.
- [16] B. Ewen-Campen, S. Donoughe, D. N. Clarke, and C. G. Extavour. Germ cell specification requires zygotic mechanisms rather than germ plasm in a basally branching insect. *Curr. Biol.*, 23(10):835–842, May 2013.
- [17] K. Illmensee and A. P. Mahowald. Transplantation of posterior polar plasm in *Drosophila*: Induction of germ cells at the anterior pole of the egg. *Proc. Natl. Acad. Sci. U. S. A.*, 71(4):1016–1020, April 1974.
- [18] R. P. Brendza, L. R. Serbus, J. B. Duffy, and W. M. Saxton. A function for kinesin I in the posterior transport of *oskar* mRNA and Stauf protein. *Science*, 289(5487):2120–2122, September 2000.
- [19] J. Kim-Ha, P. J. Webster, J. L. Smith, and P. M. Macdonald. Multiple RNA regulatory elements mediate distinct steps in localization of *oskar* mRNA. *Development*, 119(1):169–178, September 1993.
- [20] J. Kim-Ha, J. L. Smith, and P. M. Macdonald. *oskar* mRNA is localized to the posterior pole of the *Drosophila* oocyte. *Cell*, 66(1):23–35, July 1991.
- [21] F. H. Markussen, A. M. Michon, W. Breitwieser, and A. Ephrussi. Translational control of *oskar* generates short OSK, the isoform that induces pole plasm assembly. *Development*, 121(11):3723–3732, November 1995.
- [22] N. F. Vanzo and A. Ephrussi. Oskar anchoring restricts pole plasm formation to the posterior of the *Drosophila* oocyte. *Development*, 129(15):3705–3714, August 2002.
- [23] A. Kulkarni and C. G. M. Extavour. Convergent evolution of germ granule nucleators: A hypothesis. *Stem Cell Res.*, 24:188–194, October 2017.

- [24] F. Bontems, A. Stein, F. Marlow, J. Lyautey, T. Gupta, M. C. Mullins, and R. Dosch. Bucky ball organizes germ plasm assembly in zebrafish. *Curr. Biol.*, 19(5):414–422, March 2009.
- [25] I. Kawasaki, A. Amiri, Y. Fan, N. Meyer, S. Dunkelbarger, T. Motohashi, T. Karashima, O. Bossinger, and S. Strome. The PGL family proteins associate with germ granules and function redundantly in *Caenorhabditis elegans* germline development. *Genetics*, 167(2):645–661, June 2004.
- [26] I. Kawasaki, Y. H. Shim, J. Kirchner, J. Kaminker, W. B. Wood, and S. Strome. PGL-1, a predicted RNA-binding component of germ granules, is essential for fertility in *C. elegans*. *Cell*, 94(5):635–645, September 1998.
- [27] J. A. Lynch, O. Ozüak, A. Khila, E. Abouheif, C. Desplan, and S. Roth. The phylogenetic origin of *oskar* coincided with the origin of maternally provisioned germ plasm and pole cells at the base of the Holometabola. *PLoS Genet.*, 7(4):e1002029, April 2011.
- [28] J. A. Nelson. *The embryology of the honey bee*. Princeton University Press, 1915.
- [29] L. Nagy, L. Riddiford, and K. Kiguchi. Morphogenesis in the early embryo of the lepidopteran *Bombyx mori*. *Dev. Biol.*, 165(1):137–151, September 1994.
- [30] H. Kaessmann. Origins, evolution, and phenotypic impact of new genes. *Genome Res.*, 20(10):1313–1326, October 2010.
- [31] H. Innan and F. Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.*, 11(2):97–108, February 2010.
- [32] F. Jacob. Evolution and tinkering. *Science*, 196(4295):1161–1166, June 1977.
- [33] D. Tautz and T. Domazet-Lošo. The evolutionary origin of orphan genes. *Nat. Rev. Genet.*, 12(10):692–702, August 2011.
- [34] J. Cai, R. Zhao, H. Jiang, and W. Wang. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics*, 179(1):487–496, May 2008.
- [35] T. J. A. J. Heinen, F. Staubach, D. Häming, and D. Tautz. Emergence of a new gene from an intergenic region. *Curr. Biol.*, 19(18):1527–1531, September 2009.
- [36] D. G. Knowles and A. McLysaght. Recent de novo origin of human protein-coding genes. *Genome Res.*, 19(10):1752–1759, October 2009.
- [37] D. Li, Y. Dong, Y. Jiang, H. Jiang, J. Cai, and W. Wang. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.*, 20(4):408–420, April 2010.

- [38] C.-Y. Li, Y. Zhang, Z. Wang, Y. Zhang, C. Cao, P.-W. Zhang, S.-J. Lu, X.-M. Li, Q. Yu, X. Zheng, Q. Du, G. R. Uhl, Q.-R. Liu, and L. Wei. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput. Biol.*, 6(3):e1000734, March 2010.
- [39] P. J. Keeling and J. D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.*, 9(8):605–618, August 2008.
- [40] M. Shelomi, E. G. J. Danchin, D. Heckel, B. Wipfler, S. Bradler, X. Zhou, and Y. Pauchet. Horizontal Gene Transfer of Pectinases from Bacteria Preceded the Diversification of Stick and Leaf Insects. *Scientific Reports*, 6(1), 2016.
- [41] L. Boto. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc. R. Soc. B.*, 281(1777):20132450, 2014.
- [42] D. G. Quispe-Huamanquispe, G. Gheysen, and J. F. Kreuze. Horizontal Gene Transfer Contributes to Plant Evolution: The Case of Agrobacterium T-DNAs. *Frontiers in Plant Science*, 8, 2017.
- [43] N. Wybouw, Y. Pauchet, D. G. Heckel, and T. Van Leeuwen. Horizontal Gene Transfer Contributes to the Evolution of Arthropod Herbivory. *Genome Biology and Evolution*, 8(6):1785–1801, 2016.
- [44] Q. Zhou, G. Zhang, Y. Zhang, S. Xu, R. Zhao, Z. Zhan, X. Li, Y. Ding, S. Yang, and W. Wang. On the origin of new genes in *Drosophila*. *Genome Res.*, 18(9):1446–1455, September 2008.
- [45] Y. E. Zhang, M. D. Vibranovski, B. H. Krinsky, and M. Long. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. *Genome Research*, 20(11):1526–1533, 2010.
- [46] M. W. Hahn, M. V. Han, and S.-G. Han. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet.*, 3(11):e197, November 2007.
- [47] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, November 2000.
- [48] B. Lewin, J. E. Krebs, E. S. Goldstein, and S. T. Kilpatrick. *Lewin's Essential GENES*. Jones & Bartlett Publishers, March 2009.
- [49] S. Chen, Y. E. Zhang, and M. Long. New Genes in *Drosophila* Quickly Become Essential. *Science*, 330(6011):1682–1685, 2010.
- [50] W. Zhang, P. Landback, A. R. Gschwend, B. Shen, and M. Long. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.*, 16:202, October 2015.
- [51] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.

- [52] Y. E. Zhang and M. Long. New genes contribute to genetic and phenotypic novelties in human evolution. *Current Opinion in Genetics & Development*, 29:90–96, 2014.
- [53] S. Chen, M. Spletter, X. Ni, K. P. White, L. Luo, and M. Long. Frequent recent origination of brain genes shaped the evolution of foraging behavior in *Drosophila*. *Cell Rep.*, 1(2):118–132, February 2012.
- [54] M. Jeske, M. Bordi, S. Glatt, S. Müller, V. Rybin, C. W. Müller, and A. Ephrussi. The Crystal Structure of the *Drosophila* Germline Inducer Oskar Identifies Two Domains with Distinct Vasa Helicase- and RNA-Binding Activities. *Cell Reports*, 12(4):587–598, 2015.
- [55] N. Yang, Z. Yu, M. Hu, M. Wang, R. Lehmann, and R.-M. Xu. Structure of *Drosophila* Oskar reveals a novel RNA binding protein. *Proc. Natl. Acad. Sci. U. S. A.*, 112(37):11541–11546, September 2015.
- [56] M. Jeske, C. W. Müller, and A. Ephrussi. The LOTUS domain is a conserved DEAD-box RNA helicase regulator essential for the recruitment of Vasa to the germ plasm and nuage. *Genes Dev.*, 31(9):939–952, May 2017.
- [57] T. A. M. Kostas Bourtzis. *Insect Symbiosis, Volume 3*. CRC Press, 2008. doi: 10.1201/9781420064117.
- [58] A. A. Hoffmann. Essential but unhelpful wasp *Wolbachia*. *Heredity*, 103(3):194–195, 2009.
- [59] J. C. Dunning Hotopp, M. E. Clark, D. C. S. G. Oliveira, J. M. Foster, P. Fischer, M. C. Muñoz Torres, J. D. Giebel, N. Kumar, N. Ishmael, S. Wang, J. Ingram, R. V. Nene, J. Shepard, J. Tomkins, S. Richards, D. J. Spiro, E. Ghedin, B. E. Slatko, H. Tettelin, and J. H. Werren. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, 317(5845):1753–1756, September 2007.
- [60] K. M. Oliver, P. H. Degnan, G. R. Burke, and N. A. Moran. Facultative Symbionts in Aphids and the Horizontal Transfer of Ecologically Important Traits. *Annual Review of Entomology*, 55(1):247–266, 2010.
- [61] M. Ravenhall, N. Škunca, F. Lassalle, and C. Dessimoz. Inferring horizontal gene transfer. *PLoS Comput. Biol.*, 11(5):e1004095, May 2015.
- [62] B. Ewen-Campen, J. R. Srouji, E. E. Schwager, and C. G. Extavour. *oskar* predates the evolution of germ plasm in insects. *Curr. Biol.*, 22(23):2278–2283, December 2012.
- [63] S. Donoughe, T. Nakamura, B. Ewen-Campen, D. A. Green, 2nd, L. Henderson, and C. G. M. Extavour. BMP signaling is required for the generation of primordial germ cells in an insect. *Proc. Natl. Acad. Sci. U. S. A.*, 111(11):4133–4138, March 2014.

- [64] W. Breitwieser, F. H. Markussen, H. Horstmann, and A. Ephrussi. Oskar protein interaction with Vasa represents an essential step in polar granule assembly. *Genes & Development*, 10(17):2179–2188, 1996.
- [65] T. Noce, S. Okamoto-Ito, and N. Tsunekawa. Vasa homolog genes in mammalian germ cell development. *Cell Struct. Funct.*, 26(3):131–136, June 2001.
- [66] E. Raz. The function and regulation of *vasa*-like genes in germ-cell development. *Genome Biol.*, 1(3):REVIEWS1017, September 2000.
- [67] J. Anne. Targeting and Anchoring Tudor in the Pole Plasm of the *Drosophila* Oocyte. *PLoS ONE*, 5(12):e14362, 2010.
- [68] A. Ephrussi, L. K. Dickinson, and R. Lehmann. *Oskar* organizes the germ plasm and directs localization of the posterior determinant *nanos*. *Cell*, 66(1):37–50, July 1991.
- [69] I. Callebaut and J.-P. Mornon. LOTUS, a new domain associated with small RNA pathways in the germline. *Bioinformatics*, 26(9):1140–1144, 2010.
- [70] V. Anantharaman, D. Zhang, and L. Aravind. OST-HTH: a novel predicted RNA-binding domain. *Biology Direct*, 5(1):13, 2010.
- [71] K. Howard, M. Jaglarz, N. Zhang, J. Shah, and R. Warrior. Migration of *Drosophila* germ cells: analysis using enhancer trap lines. *Development*, 119(Supplement):213–218, December 1993.
- [72] G. Callaini, M. G. Riparbelli, and R. Dallai. Pole Cell Migration through the Gut Wall of the *Drosophila* Embryo: Analysis of Cell Interactions. *Developmental Biology*, 170(2):365–375, 1995.
- [73] D. E. Wheeler. Chapter 188 - Ovarioles. In V. H. Resh and R. T. Cardé, editors, *Encyclopedia of Insects (Second Edition)*, pages 743–744. Academic Press, San Diego, January 2009.
- [74] R. C. King. *Ovarian Development in Drosophila melanogaster*. Academic Press, New York, 1970.
- [75] S. H. Church, B. A. S. de Medeiros, S. Donoughe, N. L. M. Reyes, and C. G. Extavour. Repeated loss of variation in insect ovary morphology highlights the role of developmental constraint in life-history evolution. *bioRxiv*, 2020. doi: 10.1101/2020.07.07.191940. URL <https://www.biorxiv.org/content/early/2020/07/07/2020.07.07.191940>.
- [76] J. Büning. The trophic tissue of telotrophic ovarioles in polyphage coleoptera. *Zoomorphologie*, 93(1):33–50, 1979.
- [77] P. O. Ritcher, R. Po, and B. Cw. Ovariole numbers in Scarabaeoidea (Coleoptera: Lucanidae, Passalidae, Scarabaeidae). *Proceedings of the Entomological Society of Washington*, 1974.

- [78] J. Hodin. She shapes events as they come. In D. Whitman and T. N. Ananthakrishnan, editors, *Plasticity in female insect reproduction: Mechanisms and Consequences*, pages 423–521. Taylor & Francis, 2009.
- [79] R. B. R. Azevedo, V. French, and L. Partridge. Thermal evolution of egg size in *Drosophila melanogaster*. *Evolution*, 50(6):2338–2345, December 1996.
- [80] P. Capy, E. Pla, and J. R. David. Phenotypic and genetic variability of morphometrical traits in natural populations of *Drosophila melanogaster* and *D. simulans*. II. Within-population variability. *Genet. Sel. Evol.*, 26(1):15, February 1994.
- [81] J. R. David and C. Bocquet. Similarities and differences in latitudinal adaptation of two *Drosophila* sibling species. *Nature*, 257(5527):588–590, October 1975.
- [82] P. Capy, E. Pla, and J. R. David. Phenotypic and genetic variability of morphometrical traits in natural populations of *Drosophila melanogaster* and *D. simulans*. I. Geographic variations. *Genet. Sel. Evol.*, 25(6):517, December 1993.
- [83] A. S. Lobell, R. R. Kaspari, Y. L. Serrano Negron, and S. T. Harbison. The Genetic Architecture of Ovariole Number in *Drosophila melanogaster*: Genes with Major, Quantitative, and Pleiotropic Effects. *G3*, 7(7):2391–2403, July 2017.
- [84] D. P. Sarikaya, A. A. Belay, A. Ahuja, A. Dorta, D. A. Green, 2nd, and C. G. Extavour. The roles of cell size and cell number in determining ovariole number in *Drosophila*. *Dev. Biol.*, 363(1):279–289, March 2012.
- [85] A. Sebé-Pedrós, Y. Zheng, I. Ruiz-Trillo, and D. Pan. Premetazoan origin of the *hippo* signaling pathway. *Cell Rep.*, 1(1):13–20, January 2012.
- [86] J. Dong, G. Feldmann, J. Huang, S. Wu, N. Zhang, S. A. Comerford, M. F. Gayyed, R. A. Anders, A. Maitra, and D. Pan. Elucidation of a universal size-control mechanism in *Drosophila* and mammals. *Cell*, 130(6):1120–1133, September 2007.
- [87] D. P. Sarikaya and C. G. Extavour. The Hippo pathway regulates homeostatic growth of stem cell niche precursors in the *Drosophila* ovary. *PLoS Genet.*, 11(2):e1004962, February 2015.
- [88] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, April 2004.
- [89] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.-H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. J.

- Pan, M. J. Schilstra, P. J. C. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295(5560):1669–1678, March 2002.
- [90] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, August 2002.
- [91] A. Mbodj, G. Junion, C. Brun, E. E. M. Furlong, and D. Thieffry. Logical modelling of *Drosophila* signalling pathways. *Mol. Biosyst.*, 9(9):2248–2258, September 2013.
- [92] R.-S. Wang and J. Loscalzo. Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications. *J. Mol. Biol.*, 430(18 Pt A):2939–2950, September 2018.
- [93] E. Guney, J. Menche, M. Vidal, and A.-L. Barabási. Network-based in silico drug efficacy screening. *Nat. Commun.*, 7:10331, February 2016.
- [94] S. Valverde. Breakdown of Modularity in Complex Networks. *Front. Physiol.*, 8:497, July 2017.
- [95] S. D. Ghiassian, J. Menche, and A.-L. Barabási. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.*, 11(4):e1004120, April 2015.
- [96] R.-S. Wang, K. T. Hall, F. Giulianini, D. Passow, T. J. Kaptchuk, and J. Loscalzo. Network analysis of the genomic basis of the placebo effect. *JCI Insight*, 2(11):93911, June 2017.
- [97] A.-L. Barabási. *Network Science*. Cambridge University Press, July 2016.
- [98] M.-S. Kim, J.-R. Kim, D. Kim, A. D. Lander, and K.-H. Cho. Spatiotemporal network motif reveals the biological traits of developmental gene regulatory networks in *Drosophila melanogaster*. *BMC Syst. Biol.*, 6:31, May 2012.
- [99] B. Verd, N. A. Monk, and J. Jaeger. Modularity, criticality, and evolvability of a developmental gene regulatory network. *Elife*, 8:e42832, June 2019.
- [100] M. Thiel and G. Wellborn. *The Natural History of the Crustacea: Life Histories, Volume 5*. Oxford University Press, May 2018.
- [101] C. Wolff and M. Gerberding. “Crustacea”: Comparative Aspects of Early Development. In A. Wanninger, editor, *Evolutionary Developmental Biology of Invertebrates 4*, pages 63–100. Springer Wien, Vienna, 2015.
- [102] G. Scholtz and C. Wolff. Arthropod Embryology: Cleavage and Germ Band Development. In A. Minelli, G. Boxshall, and G. Fusco, editors, *Arthropod Biology and Evolution: Molecules, Development, Morphology*, pages 63–89. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

- [103] P. Nestorov, F. Battke, M. P. Levesque, and M. Gerberding. The maternal transcriptome of the crustacean *Parhyale hawaiiensis* is inherited asymmetrically to invariant cell lineages of the ectoderm and mesoderm. *PLoS One*, 8(2):e56049, February 2013.
- [104] C. Grobben. Die Entwicklungsgeschichte der *Moina rectoris*. *Arb. zool. Inst. Wien*, 2:203–268, 1879.
- [105] M. Gerberding and G. Scholtz. Cell lineage of the midline cells in the amphipod crustacean *Orchestiacavimana* (Crustacea, Malacostraca) during formation and separation of the germ band. *Dev. Genes Evol.*, 209(2):91–102, 1999.
- [106] M. Gerberding, W. E. Browne, and N. H. Patel. Cell lineage analysis of the amphipod crustacean *Parhyale hawaiiensis* reveals an early restriction of cell fates. *Development*, 129(24):5789–5801, December 2002.
- [107] S. F. Gilbert. Early Drosophila Development. In S. F. Gilbert, editor, *Developmental Biology*. 6th edition. Sinauer Associates, 2000.
- [108] T. Hashimshony, F. Wagner, N. Sher, and I. Yanai. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, 2(3):666–673, September 2012.
- [109] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.
- [110] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015.
- [111] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.-B. Fan, P. Lönnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21(7):1160–1167, July 2011.
- [112] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, January 2017.

- [113] E. Denisenko, B. B. Guo, M. Jones, R. Hou, L. de Kock, T. Lassmann, D. Poppe, O. Clément, R. K. Simmons, R. Lister, and A. R. R. Forrest. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.*, 21(1):130, June 2020.
- [114] M. Gerberding and N. H. Patel. Gastrulation in Crustacean: Germ Layers and Cell Lineages. In C. D. Stern, editor, *Gastrulation: From Cells to Embryo*, pages 79–90. CSHL Press, 2004.
- [115] F. Alwes and G. Scholtz. Cleavage and gastrulation of the euphausiacean *Meganyctiphanes norvegica* (Crustacea, Malacostraca). *Zoomorphology*, 123(3):125–137, 2004.
- [116] P. L. Hertzler. Cleavage and gastrulation in the shrimp *Penaeus (Litopenaeus) vannamei* (Malacostraca, Decapoda, Dendrobranchiata). *Arthropod Struct. Dev.*, 34(4):455–469, October 2005.
- [117] P. L. Hertzler. Development of the mesendoderm in the dendrobranchiate shrimp *Sicyonia ingentis*. *Arthropod Struct. Dev.*, 31(1):33–49, September 2002.
- [118] C. Biffis, F. Alwes, and G. Scholtz. Cleavage and gastrulation of the dendrobranchiate shrimp *Penaeus monodon* (Crustacea, Malacostraca, Decapoda). *Arthropod Struct. Dev.*, 38(6):527–540, November 2009.
- [119] C. Wolff and G. Scholtz. Cell lineage, axis formation, and the origin of germ layers in the amphipod crustacean *Orchestia cavimana*. *Dev. Biol.*, 250(1):44–58, October 2002.
- [120] J. B. Pawlak, M. J. Sellars, A. Wood, and P. L. Hertzler. Cleavage and gastrulation in the Kuruma shrimp *Penaeus (Marsupenaeus) japonicus* (Bate): A revised cell lineage and identification of a presumptive germ cell marker. *Development, Growth & Differentiation*, 52(8):677–692, 2010.
- [121] G. Scholtz and C. Wolff. Cleavage, gastrulation, and germ disc formation of the amphipod *Orchestia cavimana* (Crustacea, Malacostraca, Peracarida). *Contrib. Zool.*, 71(1-3):9–28, January 2002.
- [122] F. Alwes, B. Hitchen, and C. G. M. Extavour. Patterns of cell lineage, movement, and migration from germ layer specification to gastrulation in the amphipod crustacean *Parhyale hawaiiensis*. *Dev. Biol.*, 359(1):110–123, November 2011.
- [123] E. Taube. Beiträge zur Entwicklungsgeschichte der Euphausiden. I Die Furchung des Eis bis zur Gastrulation. *Zeitschrift für wissenschaftliche Zoologie*, 92:427–464, 1909.
- [124] E. Taube. Beiträge zur Entwicklungsgeschichte der Euphausiden. II Von der Gastrulation bis zum Furciliastadium. *Zeitschrift für wissenschaftliche Zoologie*, 94:577–658, 1915.
- [125] P. L. Hertzler and W. H. Clark, Jr. Cleavage and gastrulation in the shrimp *Sicyonia ingentis*: invagination is accompanied by oriented cell division. *Development*, 116(1):127–140, September 1992.

- [126] R. C. Chaw and N. H. Patel. Independent migration of cell populations in the early gastrulation of the amphipod crustacean *Parhyale hawaiiensis*. *Dev. Biol.*, 371(1):94–109, November 2012.
- [127] V. S. Hunnekuhl and C. Wolff. Reconstruction of cell lineage and spatiotemporal pattern formation of the mesoderm in the amphipod crustacean *Orchestia cavimana*. *Dev. Dyn.*, 241(4):697–717, April 2012.
- [128] W. E. Browne, A. L. Price, M. Gerberding, and N. H. Patel. Stages of embryonic development in the amphipod crustacean, *Parhyale hawaiiensis*. *Genesis*, 42(3):124–149, July 2005.
- [129] J. Huisken, J. Swoger, F. Del Bene, J. Wittbrodt, and E. H. K. Stelzer. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science*, 305(5686):1007–1009, August 2004.
- [130] P. J. Keller, A. D. Schmidt, A. Santella, K. Khairy, Z. Bao, J. Wittbrodt, and E. H. K. Stelzer. Fast, high-contrast imaging of animal development with scanned light sheet-based structured-illumination microscopy. *Nat. Methods*, 7(8):637–642, August 2010.
- [131] A. L. Price, M. S. Modrell, R. L. Hannibal, and N. H. Patel. Mesoderm and ectoderm lineages in the crustacean *Parhyale hawaiiensis* display intra-germ layer compensation. *Dev. Biol.*, 341(1):256–266, May 2010.
- [132] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart. The CAVE: audio visual experience automatic virtual environment. *Commun. ACM*, 35(6):64–72, June 1992.
- [133] R. P. Theart, B. Loos, and T. R. Niesler. Virtual reality assisted microscopy data visualization and colocalization analysis. *BMC Bioinformatics*, 18(S2), 2017.
- [134] C. Wolff, J.-Y. Tinevez, T. Pietzsch, E. Stamatakis, B. Harich, L. Guignard, S. Preibisch, S. Shorte, P. J. Keller, P. Tomancak, and A. Pavlopoulos. Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. *Elife*, 7:e34410, March 2018.

*Life did not take over the globe by combat, but by net-
working.*

Lynn Margulis

1

Bacterial contribution to the genesis of the novel
germ line determinant oskar

ABSTRACT

New cellular functions and developmental processes can evolve by modifying existing genes or creating novel genes. Novel genes can arise not only via duplication or mutation but also by acquiring foreign DNA, also called horizontal gene transfer (HGT). Here we show that HGT likely contributed to the creation of a novel gene indispensable for reproduction in some insects. Long considered a novel gene with unknown origin, *oskar* has evolved to fulfil a crucial role in insect germ cell formation. Our analysis of over 100 insect Oskar sequences suggests that Oskar arose *de novo* via fusion of eukaryotic and prokaryotic sequences. This work shows that highly unusual gene origin processes can give rise to novel genes that can facilitate evolution of novel developmental mechanisms.

CONTRIBUTIONS AND PUBLICATION

This work was published in the journal *Elife* in 2020 and has been reformatted as the chapter presented here:

Blondel, L.*, Jones, T. E., & Extavour, C. G. (2020). **Bacterial contribution to genesis of the novel germ line determinant *oskar***. *ELife*, 9. DOI:[10.7554/elife.45539](https://doi.org/10.7554/elife.45539)

Tamsin EM Jones collected the first collection of 46 *oskar* ortholog sequences. Cassandra G. Extavour proposed the hypothesis and directed the study design, funding, writing and reviewing. I generated the rest of the original work presented in this chapter.

1.1 Introduction

Heritable variation is the raw material of evolutionary change. Genetic variation can arise from mutation and gene duplication of existing genes (reviewed by Taylor and Raes¹), or through *de novo* processes², but the extent to which such novel, or "orphan" genes participate significantly in the evolutionary process is unclear. Mutation of existing cis-regulatory³ or protein coding regions⁴ can drive evolutionary change in developmental processes. However, recent studies in animals and fungi suggest that novel genes can also drive phenotypic change⁵. Although counterintuitive, novel genes may be integrating continuously into otherwise conserved gene networks, with a higher rate of partner acquisition than subtler variations on preexisting genes⁶. Moreover, in humans and fruit flies, a large proportion of novel genes are expressed in the brain, suggesting their participation in the evolution of major organ systems^{7,8}. However, while next generation sequencing has improved their discovery, the developmental and evolutionary significance of novel genes remains understudied.

The mechanism of formation of a novel gene may have implications for its function. Novel genes that arise by duplication, thus possessing the same biophysical properties as their parent genes, have innate potential to participate in preexisting cellular and molecular mechanisms (reviewed by Taylor and Raes¹). However, orphan genes lacking sequence similarity to existing genes must form novel functional molecular relationships with extant genes, in order to persist in the genome. When such genes arise by introduction of foreign DNA into a host genome through horizontal gene transfer (HGT), they may introduce novel, already functional sequence information into a genome. Whether genes created by HGT show a greater propensity to contribute to or enable novel processes is unclear. Endosymbionts in the host germ line cytoplasm (germ line symbionts) could increase the occurrence of evolutionarily relevant HGT events, as foreign DNA integrated into the germ line genome is transferred to the next generation. HGT from bacterial endosymbionts into insect genomes appears widespread, involving transfer of metabolic genes or even larger genomic fragments to the host genome (see for example^{9,10,11,12}).

Here we examined the evolutionary origins of the *oskar* (*osk*) gene, long considered a novel gene

that evolved to be indispensable for insect reproduction¹³. First discovered in *Drosophila melanogaster*¹⁴, *osk* is necessary and sufficient for assembly of germ plasm, a cytoplasmic determinant that specifies the germ line in the embryo. Germ plasm-based germ line specification appears derived within insects, confined to insects that undergo metamorphosis (Holometabola)^{15,16}. Initially thought exclusive to Diptera (flies and mosquitoes), its discovery in a wasp, another holometabolous insect with germ plasm¹⁷, led to the hypothesis that *oskar* originated as a novel gene at the base of the Holometabola approximately 300 Mya, facilitating the evolution of insect germ plasm as a novel developmental mechanism¹⁷. However, its subsequent discovery in a cricket¹⁵, a hemimetabolous insect without germ plasm¹⁸, implied that *osk* was instead at least 50 My older, and that its germ plasm role was derived rather than ancestral¹⁹. Despite its orphan gene status, *osk* plays major developmental roles, interacting with the products of many genes highly conserved across animals^{20,21,22}. *osk* thus represents an example of a novel gene that not only functions within pre-existing gene networks in the nervous system¹⁵, but has also evolved into the only animal gene that has been experimentally demonstrated to be both necessary and sufficient to specify functional primordial germ line cells^{23,24}.

1.2 Results

The evolutionary origins of this remarkable gene are unknown. *Osk* contains two biophysically conserved domains, an N-terminal LOTUS domain and a C-terminal hydrolase-like domain called OSK^{21,25} (Figure 1.1a). An initial BLASTp search using the full-length *D. melanogaster osk* sequence as a query yielded either other holometabolous insect *osk* genes, or partial hits for the LOTUS or OSK domains (E-value < 0.01; **Supplementary files**: BLAST search results). This suggested that full length *osk* was unlikely to be a duplication of any other known gene. This prompted us to perform two more BLASTp searches, one using each of the two conserved *Osk* protein domains individually as query sequences. Strikingly, in this BLASTp search, although we recovered several eukaryotic hits for the LOTUS domain, we recovered no eukaryotic sequences that resembled the OSK domain, even with very low E-value stringency (E-value < 10; see Methods section "*BLAST searches of oskar*" for an explanation of E-value threshold choices;

Supplementary files: BLAST search results).

To understand this anomaly, we built an alignment of 95 Oskar sequences (**Supplementary files:** Alignments>OSKAR_MUSCLE_FINAL.fasta; Tables A.1 and A.2) and used a custom iterative HMMER sliding window search tool to compare each domain with protein sequences from all domains of life. Sequences most similar to the LOTUS domain were almost exclusively eukaryotic sequences (Table A.3). In contrast, those most similar to the OSK domain were bacterial, specifically sequences similar to SGNH-like hydrolases^{21,25} (Pfam Clan: SGNH_hydrolase - CL0264; Table A.4; Figure 1.1b). To visualize their relationships, we graphed the sequence similarity network for the sequences of these domains and their closest hits. We observed that the majority of LOTUS domain sequences clustered within eukaryotic sequences (Figure 1.1c). In contrast, OSK domain sequences formed an isolated cluster, a small subset of which formed a connection to bacterial sequences (Figure 1.1d). These data are consistent with a previous suggestion, based on BLAST results¹⁷, that HGT from a bacterium into an ancestral insect genome may have contributed to the evolution of *osk*. However, this possibility was not formally addressed by previous analyses, which were based on alignments of full length *Osk* containing only eukaryotic sequences as outgroups¹⁵. To rigorously test this hypothesis, we therefore performed phylogenetic analyses of the two domains independently. A finding that LOTUS sequences were nested within eukaryotes, while OSK sequences were nested within bacteria, would provide support for the HGT hypothesis.

Both Maximum likelihood and Bayesian approaches confirmed this prediction (Figure 1.2a, Figure A.1, Figure A.2), and these results were robust to changes in the methods of sequence alignment (Figure A.1). As expected, LOTUS sequences from *Osk* proteins were related to other eukaryotic LOTUS domains, to the exclusion of the only three bacterial sequences that met our E-value cutoff for inclusion in the analyses (Figures A.1 to A.2; see Methods). LOTUS sequences from non-*Oskar* proteins were almost exclusively eukaryotic. (Table A.3); only three bacterial sequences matched the LOTUS domain with an E-value < 0.01. *Osk* LOTUS domains clustered into two distinct clades, one comprising all Dipteran sequences, and the other comprising all other *Osk* LOTUS domains examined from both holometabolous and hemimetabolous orders (Figure 1.2a). Dipteran *Osk* LOTUS sequences formed a monophyletic group that branched sister

to a clade of LOTUS domains from Tud5 family proteins of non-arthropod animals (NAA). NAA LOTUS domains from Tud7 family members were polyphyletic, but most of them formed a clade branching sister to (Osk LOTUS + NAA Tud5 LOTUS). Non-Dipteran Osk LOTUS domains formed a monophyletic group that was related in a polytomy to the aforementioned (NAA Tud7 LOTUS + (Dipteran Osk LOTUS + NAA Tud5 LOTUS)) clade, and to various arthropod Tud7 family LOTUS domains.

The fact that Tud7 LOTUS domains are polyphyletic suggests that arthropod domains in this family may have evolved differently than their homologues in other animals. The relationships of Dipteran LOTUS sequences were consistent with the current hypothesis for interrelationships between Dipteran species²⁶. Similarly, among the non-Dipteran Osk LOTUS sequences, the hymenopteran sequences form a clade to the exclusion of the single hemimetabolous sequence (from the cricket *Gryllus bimaculatus*), consistent with the monophyly of Hymenoptera²⁷. It is unclear why Dipteran Osk LOTUS domains cluster separately from those of other insect Osk proteins. We speculate that the evolution of the Long Oskar domain^{28,29}, which appears to be a novelty within Diptera ([Supplementary Files: Alignments>OSKAR_MUSCLE_FINAL.fasta](#)), may have influenced the evolution of the Osk LOTUS domain in at least some of these insects.

Consistent with this hypothesis, of the 17 Dipteran *oskar* genes we examined, the seven *oskar* genes possessing a Long Osk domain clustered into two clades based on the sequences of their LOTUS domain. One of these clades comprised five *Drosophila* species (*D. willistoni*, *D. mojavensis*, *D. virilis*, *D. grimshawi* and *D. immigrans*), and the second was composed of two calyptrate flies from different superfamilies, *Musca domestica* (Muscoidea) and *Lucilia cuprina* (Oestroidea).

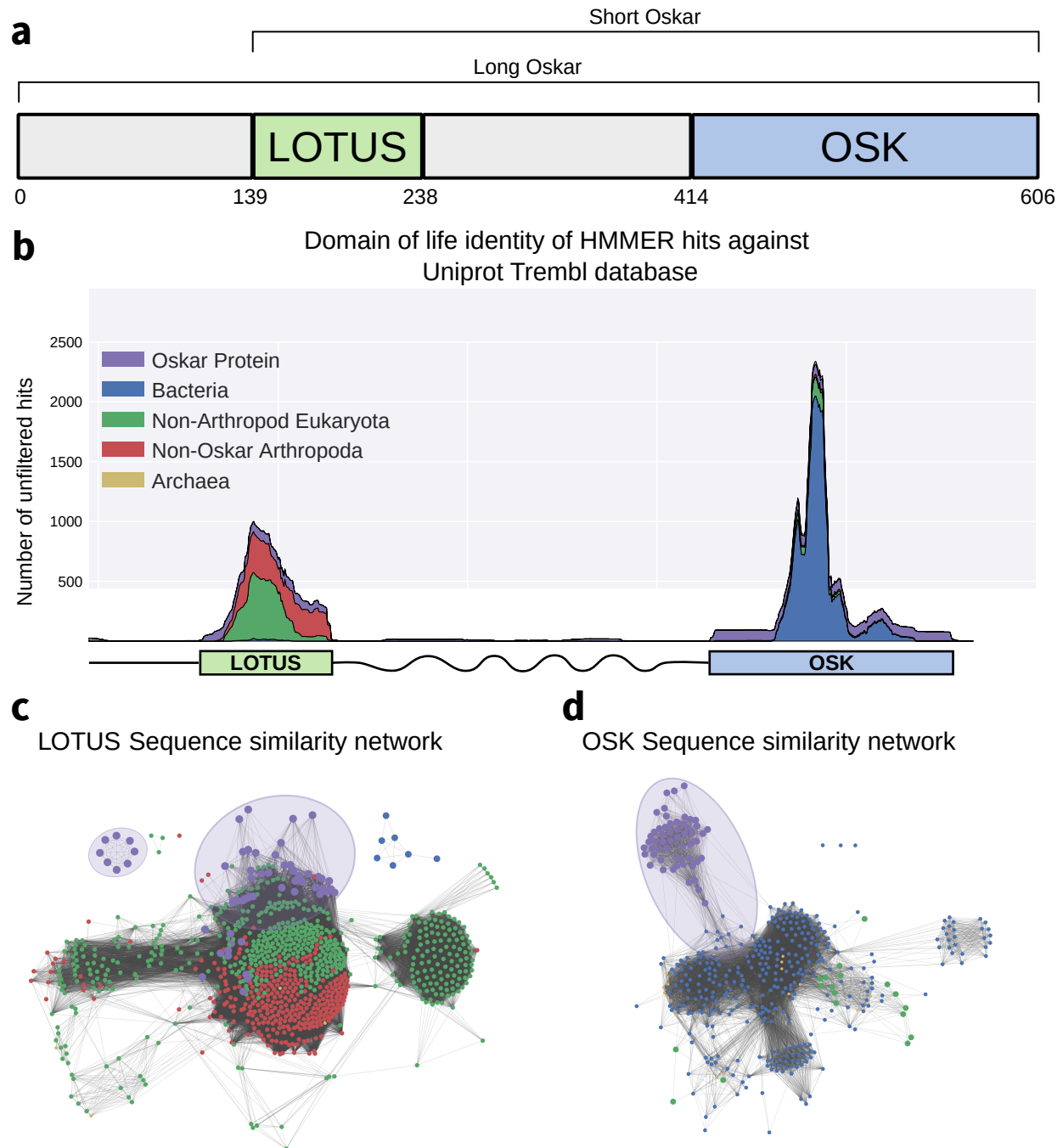
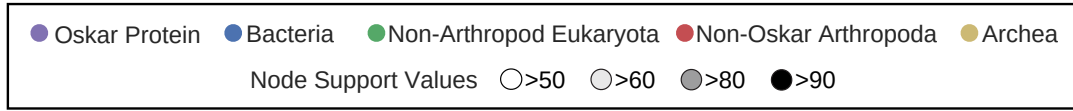


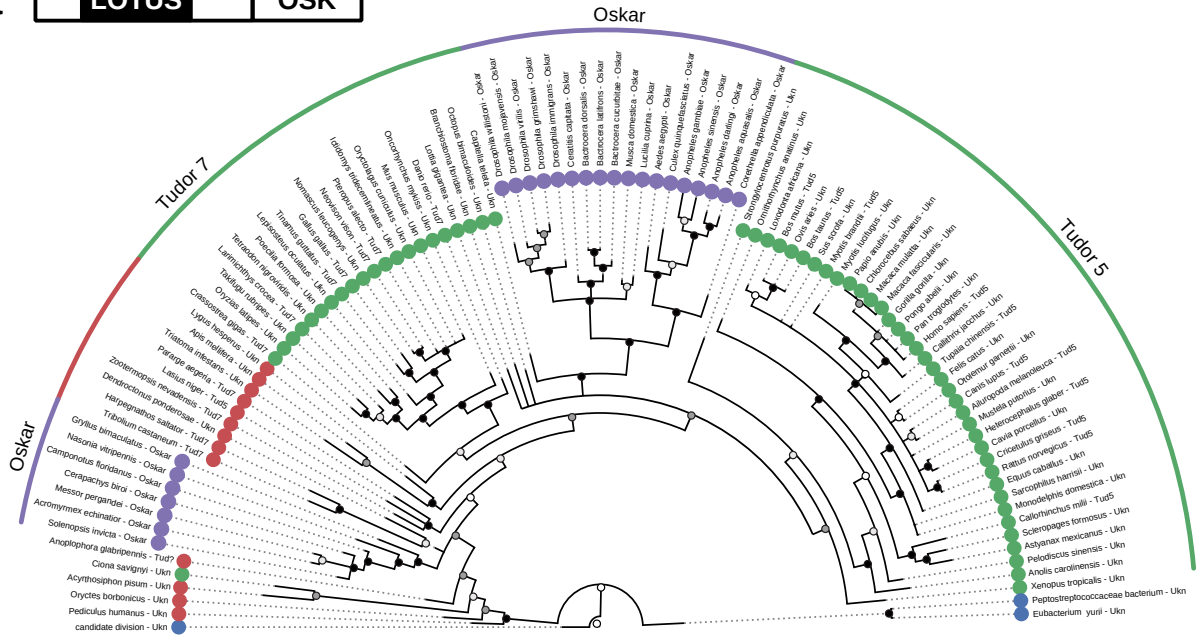
Figure 1.1: Sequence analysis of the Oskar gene. **a)** Schematic representation of the Oskar gene. The LOTUS and OSK hydrolase-like domains are separated by a poorly conserved region of predicted high disorder and variable length between species. In some dipterans, a region 5' to the LOTUS domain is translated to yield a second isoform, called Long Oskar. Residue numbers correspond to the *D. melanogaster* Osk sequence. **b)** Stackplot of domain of life identity of HMMER hits across the protein sequence. For a sliding window of 60 Amino Acids across the protein sequence (X axis), the number of hits in the Trembl (UniProt) database (Y axis) is represented and color coded by domain of life origin (see Methods: Iterative HMMER search of OSK and LOTUS domains), stacked on top of each other. **c & d)** EFI-EST-generated graphs of the sequence similarity network of the LOTUS (**c**) and OSK (**d**) domains of Oskar³⁰. Sequences were obtained using HMMER against the UniProtKB database. Most Oskar LOTUS sequences cluster within eukaryotes and arthropods. In contrast, Oskar OSK sequences cluster most strongly with a small subset of bacterial sequences.

Figure 1.2 (following page): Phylogenetic analysis of the LOTUS and OSK domains. **a)** Bayesian consensus tree for the LOTUS domain. Three major LOTUS-containing protein families are represented within the tree: Tudor 5, Tudor 7, and Oskar. Oskar LOTUS domains form two clades, one containing only dipterans and one containing all other represented insects (hymenopterans and orthopterans). The tree was rooted to the three bacterial sequences added in the dataset. **b)** Bayesian consensus tree for the OSK domain. The OSK domain is nested within GDSL-like domains of bacterial species from phyla known to contain germ line symbionts in insects. The ten non-Oskar eukaryotic sequences in the analysis form one clade comprising fungal Carbohydrate Active Enzyme 3 (CAZ3) proteins. For Bayesian and RaxML trees with all accession numbers and node support values see Figures A.1 to A.4.

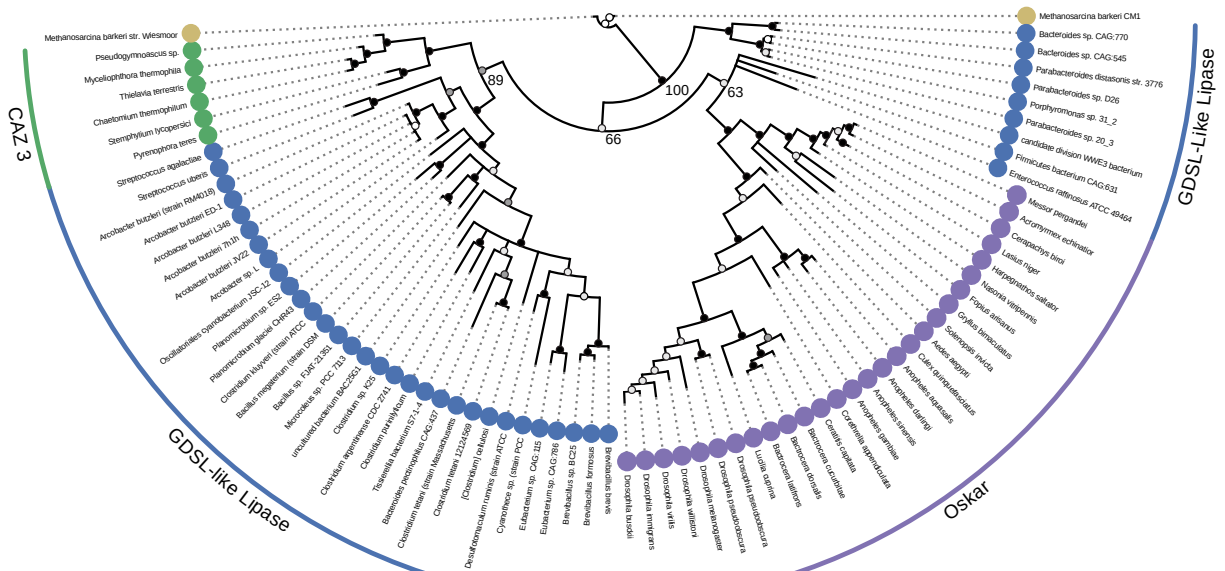
Figure 1.2: (continued)



a **LOTUS** **OSK**



b **LOTUS** **OSK**



In summary, the LOTUS domain of Osk proteins is most closely related to a number of other LOTUS domains found in eukaryotic proteins, as would be expected for a gene of animal origin, and the phylogenetic interrelationships of these sequences are largely consistent with the current species or family level trees for the corresponding insects.

In contrast, OSK domain sequences were nested within bacterial sequences (Figure 1.2b, Figures A.3 and A.4). This bacterial, rather than eukaryotic, affinity of the OSK domain was recovered even when different sequence alignment methods were used (Figures A.7 to A.11). The only eukaryotic proteins emerging from the iterative HMMER search for OSK domain sequences that had an E-value < 0.01 were all from fungi. All five of these sequences were annotated as Carbohydrate Active Enzyme 3 (CAZ3), and all CAZ3 sequences formed a clade that was sister to a clade of primarily Firmicutes. Most bacterial sequences used in this analysis were annotated as lipases and hydrolases, with a high representation of GDSL-like hydrolases (Table A.4). OSK sequences formed a monophyletic group but did not branch sister to the other eukaryotic sequences in the analysis. Within this OSK clade, the topology of sequence relationships was largely concordant with the species tree for insects³¹, as we recovered monophyletic Diptera to the exclusion of other insect species. However, the single orthopteran OSK sequence (from the cricket *Gryllus bimaculatus*) grouped within the Hymenoptera, rather than branching as sister to all other insect sequences in the tree, as would be expected for this hemimetabolous sequence³¹.

Importantly, OSK sequences did not simply form an outgroup to bacterial sequences. To formally reject the possibility that the eukaryotic OSK clade has a sister group relationship to all bacterial sequences in the analysis, we performed topology constraint analyses using the Swofford–Olsen–Waddell–Hillis (SOWH) test, which assigns statistical support to alternative phylogenetic topologies³². We used the SOWHAT tool³³ to compare the HGT-supporting topology to two alternative topologies with constraints more consistent with vertical inheritance. The first was constrained by domain of life, disallowing paraphyletic relationships between sequences from the same domain of life (Figure A.5a). The second required monophyly of Eukaryota but allowed paraphyletic relationships between bacterial and archaeal sequences (Figure A.5b). We found that the topologies of both of these constrained trees were significantly

worse than the result we had recovered with our phylogenetic analysis (Figure A.5), namely that the closest relatives of the OSK domain were bacterial rather than eukaryotic sequences (Figures 1.2 to A.4).

OSK sequences formed a well-supported clade nested within bacterial GDSL-like lipase sequences. The majority of these bacterial sequences were from the Firmicutes, a bacterial phylum known to include insect germline symbionts^{34,35}. All other sequences from classified bacterial species, including a clade branching as sister to all other sequences, belonged either to the Bacteroidetes or to the Proteobacteria. Members of both of these phyla are also known germline symbionts of insects^{9,36} and other arthropods³⁷. In sum, the distinct phylogenetic relationships of the two domains of Oskar are consistent with a bacterial origin for the OSK domain. Further, the specific bacterial clades close to OSK suggest that an ancient arthropod germ line endosymbiont could have been the source of a GDSL-like sequence that was transferred into an ancestral insect genome, and ultimately gave rise to the OSK domain of *oskar* (Figure 1.3).

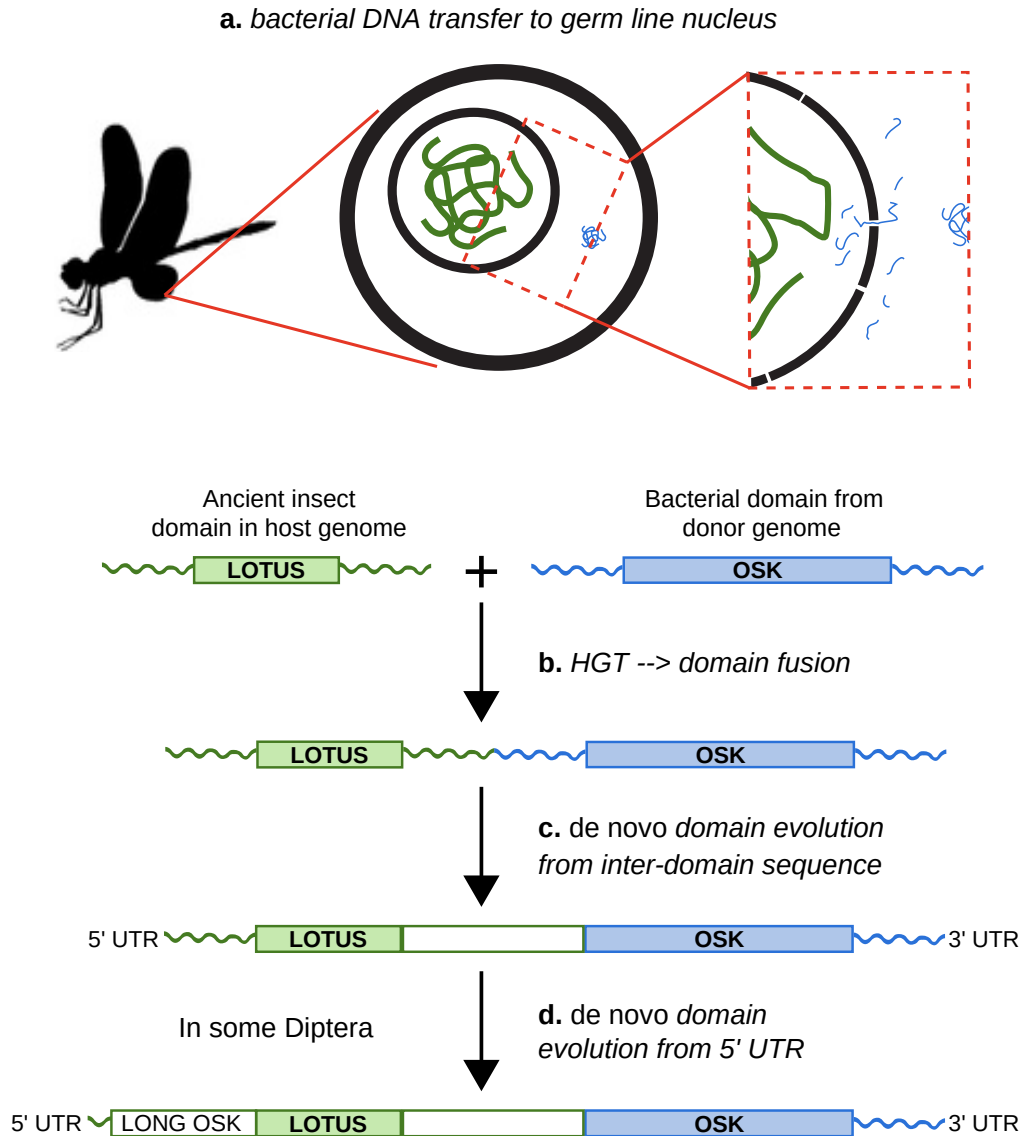


Figure 1.3: Hypothesis for the origin of *oskar*. Integration of the OSK domain close to a LOTUS domain in an ancestral insect genome. **a)** DNA containing a GDSL-like domain from an endosymbiotic germ line bacterium is transferred to the nucleus of a germ cell in an insect common ancestor. **b)** DNA damage or transposable element activity induces an integration event in the host genome, close to a pre-existing LOTUS-like domain. **c)** The region between the two domains undergoes *de novo* coding evolution, creating an open reading frame with a unique, chimeric domain structure. **d)** In some Diptera, including *D. melanogaster*, part of the 5' UTR of *oskar* has undergone *de novo* coding evolution to form the Long Oskar domain.

1.3 Discussion

While multiple mechanisms can give rise to novel genes, HGT is arguably among the least well understood, as it involves multiple genomes and ancient biotic interactions between donor and host organisms that are often difficult to reconstruct. In the case of *oskar*, however, the fact that both germline symbionts³⁸ and HGT events⁹ are widespread in insects, provides a plausible biological mechanism consistent with our hypothesis that fusion of eukaryotic and bacterial domain sequences led to the birth of this novel gene. Under this hypothesis, this fusion would have taken place before the major diversification of insects, nearly 500 million years ago³¹.

Once arisen, novel genes might be expected to disappear rapidly, given that pre-existing gene regulatory networks operated successfully without them (reviewed by Taylor and Raes¹). However, it is clear that novel genes can evolve functional connections with existing networks, become essential³⁹, and in some cases lead to new functions⁴⁰ and contribute to phenotypic diversity⁵. Even given the growing number of convincing examples of HGT from both prokaryotic and eukaryotic origins (see for example Husnik and McCutcheon⁴¹, Lelio et al.⁴², Wybouw et al.⁴³, Quispe-Huamanquispe et al.⁴⁴), some authors suspect that the contribution of horizontal gene transfer to the acquisition of novel traits has been underestimated across animals⁴⁵. Moreover, the functional contribution of genes horizontally transferred specifically from bacteria to insects has been documented for a range of adaptive phenotypes (see for example^{46,47,48}), including digestive metabolism^{10,11,49}, glycolysis⁵⁰ complex symbiosis¹² and endosymbiont cell wall construction⁵¹. *oskar* plays multiple critical roles in insect development, from neural patterning^{15,52} to oogenesis⁵³. In the Holometabola, a clade of nearly one million extant species⁵⁴, *oskar*'s co-option to become necessary and sufficient for germ plasm assembly is likely the cell biological mechanism underlying the evolution of this derived mode of insect germ line specification^{15,17,19}. Our study thus provides evidence that HGT can not only introduce functional genes into a host genome, but also, by contributing sequences of individual domains, generate genes with entirely novel domain structures that may facilitate the evolution of novel developmental mechanisms.

1.4 Methods

1.4.1 BLAST searches of Oskar

All BLAST searches were performed using the NCBI BLASTp tool suite⁵⁵ on the non-redundant (nr) database. Amino Acid (AA) sequences of *D. melanogaster* full length Oskar (EMBL ID AAF54306.1), as well as the AA sequences for the *D. melanogaster* Oskar LOTUS (AA 139-238) and OSK (AA 414-606) domains were used for the BLAST searches. We used the default NCBI cut-off parameters (E-value cut-off of 10) for searches using OSK and LOTUS as queries, and a more stringent E-value threshold of 0.01 for the search using full length *D. melanogaster* Oskar as a query. We chose an E-value threshold of 10 for LOTUS and OSK to capture potentially highly divergent homologs of the two domains, especially for the OSK domain, where we were looking for any viable candidate for a homologous eukaryotic domain. All BLAST searches results are included in the [Supplementary files](#): BLAST search results.

1.4.2 Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains

101 1KITE transcriptomes⁵⁶ (Table A.1) were downloaded and searched using the local BLAST program (BLAST+) using the tblastn algorithm with default parameters, with Oskar protein sequences of *Drosophila melanogaster*, *Aedes aegypti*, *Nasonia vitripennis* and *Gryllus bimaculatus* as queries (EntrezIDs: NP_731295.1, ABC41128.1, NP_001234884.1 and AFV31610.1 respectively). For all of these 1KITE transcriptome searches, predicted protein sequences from transcript data were obtained by in silico translation using the online ExpASY translate tool (<https://web.expasy.org/translate/>), taking the longest open reading frame. Publicly available sequences in the non-redundant (nr), TSA databases at NCBI, and a then-unpublished transcriptome⁵⁷ (kind gift of Matthew Benton and Siegfried Roth, University of Cologne) were subsequently searched using the web-based BLAST tool hosted at NCBI, using the tblastn algorithm with default parameters. Sequences used for queries were the four Oskar proteins described above, and newfound *oskar* sequences from the 1KITE transcriptomes of *Baetis*

pumilis, *Cryptocercus wright*, and *Frankliniella cephalica*. For both searches, *oskar* orthologs were identified by the presence of BLAST hits on the same transcript to both the LOTUS (N-terminal) and OSK (C-terminal) regions of any of the query *oskar* sequences, regardless of E-values. The sequences found were aligned using MUSCLE (8 iterations)⁵⁸ into a 46-sequence alignment (Supplementary files: Alignments>OSKAR_MUSCLE_INITIAL.fasta). From this alignment, the LOTUS and OSK domains were extracted (Supplementary files: Alignments>LOTUS_MUSCLE_INITIAL.fasta and Alignments>OSK_MUSCLE_INITIAL.fasta) to define the initial Hidden Markov Models (HMM) using the hmmbuild tool from the HMMER tool suite with default parameters⁵⁹. 126 insect genomes and 128 insect transcriptomes (from the Transcriptome Shotgun Assembly TSA database: <https://www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSA>) were subsequently downloaded from NCBI (download date September 29, 2015 ; Table A.1). Genomes were submitted to Augustus v2.5.5⁶⁰ (using the *D. melanogaster* exon HMM predictor) and SNAP v2006-07-28⁶¹ (using the default 'fly' HMM) for gene discovery. The resulting nucleotide sequence database comprising all 309 downloaded and annotated genomes and transcriptomes, was then translated in six frames to generate a non-redundant amino acid database (where all sequences with the same amino acid content are merged into one). This process was automated using a series of custom scripts available here: <https://github.com/Xqua/Genomes>. The non-redundant amino acid database was searched using the HMMER v3.1 tool suite⁵⁹ and the HMM for the LOTUS and OSK domains described above. A hit was considered positive if it consisted of a contiguous sequence containing both a LOTUS domain and an OSK domain, with the two domains separated by an inter-domain sequence. We imposed no length, alignment or conservation criteria on the inter-domain sequence, as this is a rapidly-evolving region of Oskar protein with predicted high disorder^{21,25,62}. Positive hits were manually curated and added to the main alignment, and the search was performed iteratively until no more new sequences meeting the above criteria were discovered. This resulted in a total of 95 Oskar protein sequences, (see Table A.2 for the complete list). Using the final resulting alignment (Supplementary Files: Alignments>OSKAR_MUSCLE_FINAL.fasta), the LOTUS and OSK domains were extracted from these sequences (Supplementary Files: Alignments>LOTUS_MUSCLE_FINAL.fasta and Alignments>OSK_MUSCLE_FINAL.fasta), and the final three HMM (for full-length Oskar, OSK,

and LOTUS domains) used in subsequent analyses were created using hmmbuild with default parameters (**Supplementary files:** HMM>OSK.hmm, HMM>LOTUS.hmm and HMM>OSKAR.hmm).

1.4.3 Iterative HMMER search of OSK and LOTUS domains

A reduced version of TrEMBL⁶³ (v2016-06) was created by concatenating all hits (regardless of E-value) for sequences of the LOTUS domain, the OSK domain and full-length Oskar, using hmmsearch with default parameters and the HMM models created above from the final alignment. This reduced database was created to reduce potential false positive results that might result from the limited size of the sliding window used in the search approach described here. The full-length Oskar alignment of 1133 amino acids (**Supplementary files:** Alignments>OSKAR_MUSCLE_FINAL.fasta) was split into 934 sub-alignments of 60 amino acids each using a sliding window of one amino acid. Each alignment was converted into a HMM using hmmbuild, and searched against the reduced TrEMBL database using hmmsearch using default parameters. Domain of life origin of every hit sequence at each position was recorded. Eukaryotic sequences were further classified as Oskar/Non-Oskar and Arthropod/Non-Arthropod. Finally, for the whole alignment, the counts for each category were saved and plotted in a stack plot representing the proportion of sequences from each category to create Figure 1.1b. The python code used for this search is available at <https://github.com/Xqua/Iterative-HMMER>.

1.4.4 Sequence Similarity Networks

LOTUS and OSK domain sequences from the final alignment obtained as described above (see "*Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains*"; **Supplementary files:** Alignments>LOTUS_MUSCLE_FINAL.fasta and Alignments>OSK_MUSCLE_FINAL.fasta) were searched against TrEMBL⁶⁴ (v2016-06) using HMMER. All hits with E-value < 0.01 were consolidated into a fasta file that was then entered into the EFI-EST tool³⁰ using default parameters to generate a sequence similarity network. An alignment score corresponding to 30% sequence identity was chosen for the generation of the

final sequence similarity network. Finally, the network was graphed using Cytoscape 3⁶⁵.

1.4.5 Phylogenetic Analysis Based on MUSCLE Alignment

For both the LOTUS and OSK domains, in cases where more than one sequence from the same organism was retrieved by the search described above in "*Iterative HMMER Search of OSK and LOTUS domains*", only the sequence with the lowest E-value was used for phylogenetic analysis. For the LOTUS domain, the first 97 best hits (lowest E-value) were selected, and the only three bacterial sequences that satisfied an E-value < 0.01 were manually added. For *oskar* sequences, if more than one sequence per species was obtained by the search, only the single sequence per species with the lowest E-value was kept for analysis, generating a set of 100 sequences for the LOTUS domain, and 87 sequences for the OSK domain. Unique identifiers for all sequences used to generate alignments for phylogenetic analysis are available in Tables A.3 and A.4. For both datasets, the sequences were then aligned using MUSCLE⁵⁸ (8 iterations) and trimmed using trimAl⁶⁶ with 70% occupancy. The resulting alignments that were subject to phylogenetic analysis are available in **Supplementary Files**: Alignments>LOTUS_MUSCLE_TREE.fasta and Alignments>OSK_MUSCLE_TREE.fasta. For the maximum likelihood tree, we used RaxML v8.2.4⁶⁷ with 1000 bootstraps, and the models were selected using the automatic RaxML model selection tool. The substitution model chosen for both domains was LGF. For the Bayesian tree inference, we used MrBayes V3.2.6⁶⁸ with a Mixed model (prset aamodel=Mixed) and a gamma distribution (lset rates=Gamma). We ran the MonteCarlo for 4 million generations (std < 0.01) for the OSK domain, and for 3 million generations (std < 0.01) for the LOTUS domain. For the tree comparisons (Figures A.8 and A.9), the RaxML best tree output from the MUSCLE and PRANK alignments were compared using the tool Phylo.io⁶⁹.

1.4.6 Phylogenetic analysis based on PRANK alignment

For the OSK domain, the raw full length sequences obtained from the HMMER search were aligned to each other using HMMER HMM based alignment tool: hmmlalign, with the same HMM used to do the search, namely OSK.hmm (**supplementary data**: Data/HMM/OSK.hmm). Starting from this base alignment, we used the default alignment method option offered by

PRANK (version: v.170427)⁷⁰. We then used PRANK to realign those sequences, which in turn led to a usable alignment for phylogenetic analysis. This alignment was trimmed using the same parameters as described in *Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains* above. The final alignment is available in **supplementary data:** Alignment/OSK_prank_aligned.fasta. We then performed a phylogenetic analysis of this alignment using RAXML with the same parameters described in *Phylogenetic Analysis Based on MUSCLE Alignment* above. The resulting tree is presented in Figures A.7 and A.8.

For the LOTUS domain, the raw full length sequences obtained from the HMMER search were aligned to each other using the HMMER HMM based alignment tool: hmalign, with the same HMM used to do the search, namely LOTUS.hmm (**Supplementary data:** Data/HMM/LOTUS.hmm). Starting from this base alignment, we then used PRANK with default options to realign those sequences. This alignment was trimmed using the same parameters as described in the *Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains*. The final alignment is available in **supplementary data:** Alignments/LOTUS_prank_aligned.fasta. We then performed a phylogenetic analysis using RAXML with the same parameters described above in *Phylogenetic Analysis Based on MUSCLE alignment*. The resulting trees are presented in Figures A.6 and A.9.

1.4.7 Phylogenetic Analysis Based on T Coffee alignment

For the LOTUS and OSK domain, the raw full length sequences obtained from the HMMER search were aligned to each other using T-Coffee with its default parameters⁷¹. This alignment was trimmed using the same parameters as described in *Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains* above. The final alignment is available in **supplementary data:** Alignment/LOTUS_tcoffee_aligned.fasta Alignment/OSK_tcoffee_aligned.fasta. We then performed a phylogenetic analysis of this alignment using RAXML with the same parameters described in *Phylogenetic Analysis Based on MUSCLE Alignment* above. The resulting trees are presented in Figures A.10 and A.11.

1.4.8 Visual Comparison of Phylogenetic Trees

To compare the trees obtained with different alignment tools, we used Phylo.io⁶⁹. The trees were imported in Newick format, and the Phylo.io tool generated the mirrored and aligned versions of the trees represented in Figures A.8, A.9, A.12 and A.13. The color of the branches is the tree similarity score, where lighter colors represent a higher number of topological differences. Exactly, it is a custom implementation of the Jacard Index by Phylo.io.

1.4.9 Statistical Analysis of Tree Topology

To statistically evaluate our best-supported topology of the OSK and LOTUS trees, we compared constrained topologies to the highest likelihood trees using the SOWHAT tool³³. SOWHAT automates the stringent SOWH phylogenetic topology test³², and compares the log likelihood between generated trees. We defined three constrained trees to test our results, one requiring monophyly of all domains of life, a second requiring only eukaryotic monophyly, and the last one requiring monophyly of the oskar LOTUS domain (**Supplementary Files:** Data>Trees>constrained_kingdom_tree.tre, constrained_eukmono_tree.tre & constrained_lotus_mono_tree.tre). We then ran SOWHAT using its default parameters, 1000 bootstraps, and the two constrained trees against the OSK or LOTUS alignment used to generate the phylogenetic trees (**Supplementary Files:** Alignments>OSK_MUSCLE_TREE.fasta & LOTUS_MUSCLE_TREE.fasta). All best trees generated by SOWHAT are available in (**Supplementary Files:** Data>Trees>SOWHAT_*_test.tre).

1.4.10 Data Availability

All sequences discovered using the automatic annotation pipeline described in (M&M HMM and oskar search) are annotated as such in Table A.2.

The data generated and used throughout this study can be downloaded inside the github repository at https://github.com/extavourlab/Oskar_HGT

1. Subfolder **Alignments:** All sequences identified and analyzed in this study, in FASTA format and with corresponding Alignments

2. Subfolder **BLAST search results**: Results of BLASTP searches with full length Oskar, OSK or LOTUS domains as queries
3. Subfolder **Data**: Necessary files for running the different IPython notebooks:
 - a. Subfolder **HMM**: HMM models used for iterative searching for sequences similar to full-length Oskar, LOTUS and OSK domains
 - b. Subfolder **Taxonomy**: Conversion table for UniProt ID to taxon information. (uniprot_ID_taxa.tsv)
 - c. Subfolder **Trees**: Contains the tree files obtained from
 - i. RaxML phylogenetic analyses of the OSK and LOTUS domains aligned with MUSCLE, T-Coffee or PRANK
 - ii. MrBayes phylogenetic analyses of the OSK and LOTUS domains aligned with MUSCLE
 - iii. SOWHAT analyses.

1.4.11 Code Availability

All custom code generated for this study is available in the GitHub repository

https://github.com/extavourlab/Oskar_HGT

commit ID 6f6c4c50dfb9391567d70f9eea922f3876a4e153.

1.5 Acknowledgments:

We thank Sean Eddy, Chuck Davis, and Extavour lab members for discussion.

References

- [1] J. S. Taylor and J. Raes. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.*, 38:615–643, 2004.
- [2] D. Tautz and T. Domazet-Lošo. The evolutionary origin of orphan genes. *Nat. Rev. Genet.*, 12(10):692–702, August 2011.
- [3] P. J. Wittkopp and G. Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.*, 13(1):59–69, December 2011.
- [4] H. E. Hoekstra and J. A. Coyne. The locus of evolution: evo devo and the genetics of adaptation. *Evolution*, 61(5):995–1016, May 2007.
- [5] S. Chen, B. H. Krinsky, and M. Long. New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.*, 14(9):645–660, September 2013.
- [6] W. Zhang, P. Landback, A. R. Gschwend, B. Shen, and M. Long. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.*, 16:202, October 2015.
- [7] Y. E. Zhang, P. Landback, M. Vibranovski, and M. Long. New genes expressed in human brains: implications for annotating evolving genomes. *Bioessays*, 34(11):982–991, November 2012.

- [8] S. Chen, M. Spletter, X. Ni, K. P. White, L. Luo, and M. Long. Frequent recent origination of brain genes shaped the evolution of foraging behavior in *Drosophila*. *Cell Rep.*, 1(2): 118–132, February 2012.
- [9] J. C. Dunning Hotopp, M. E. Clark, D. C. S. G. Oliveira, J. M. Foster, P. Fischer, M. C. Muñoz Torres, J. D. Giebel, N. Kumar, N. Ishmael, S. Wang, J. Ingram, R. V. Nene, J. Shepard, J. Tomkins, S. Richards, D. J. Spiro, E. Ghedin, B. E. Slatko, H. Tettelin, and J. H. Werren. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, 317(5845):1753–1756, September 2007.
- [10] R. Acuña, B. E. Padilla, C. P. Flórez-Ramos, J. D. Rubio, J. C. Herrera, P. Benavides, S.-J. Lee, T. H. Yeats, A. N. Egan, J. J. Doyle, and J. K. C. Rose. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc. Natl. Acad. Sci. U. S. A.*, 109(11): 4197–4202, March 2012.
- [11] D. B. Sloan, A. Nakabachi, S. Richards, J. Qu, S. C. Murali, R. A. Gibbs, and N. A. Moran. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Mol. Biol. Evol.*, 31(4):857–871, April 2014.
- [12] F. Husnik, N. Nikoh, R. Koga, L. Ross, R. P. Duncan, M. Fujie, M. Tanaka, N. Satoh, D. Bachtrog, A. C. C. Wilson, C. D. von Dohlen, T. Fukatsu, and J. P. McCutcheon. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, 153(7):1567–1578, June 2013.
- [13] R. Lehmann. Germ plasm biogenesis—an Oskar-centric perspective. In *Current topics in developmental biology*, volume 116, pages 679–707. Elsevier, 2016.
- [14] R. Lehmann and C. Nüsslein-Volhard. Abdominal segmentation, pole cell formation, and embryonic polarity require the localized activity of *oskar*, a maternal gene in *Drosophila*. *Cell*, 47(1):141–152, October 1986.
- [15] B. Ewen-Campen, J. R. Srouji, E. E. Schwager, and C. G. Extavour. *Oskar* predates the evolution of germ plasm in insects. *Curr. Biol.*, 22(23):2278–2283, December 2012.

- [16] C. G. Extavour and M. Akam. Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development*, 130(24):5869–5884, December 2003.
- [17] J. A. Lynch, O. Ozüak, A. Khila, E. Abouheif, C. Desplan, and S. Roth. The phylogenetic origin of *oskar* coincided with the origin of maternally provisioned germ plasm and pole cells at the base of the Holometabola. *PLoS Genet.*, 7(4):e1002029, April 2011.
- [18] B. Ewen-Campen, S. Donoughe, D. N. Clarke, and C. G. Extavour. Germ cell specification requires zygotic mechanisms rather than germ plasm in a basally branching insect. *Curr. Biol.*, 23(10):835–842, May 2013.
- [19] E. Abouheif. Evolution: *oskar* reveals missing link in co-optive evolution. *Curr. Biol.*, 23(1):R24–5, January 2013.
- [20] R. Lehmann. Germ Plasm Biogenesis—An Oskar-Centric Perspective. *Curr. Top. Dev. Biol.*, 116:679–707, February 2016.
- [21] M. Jeske, M. Bordi, S. Glatt, S. Müller, V. Rybin, C. W. Müller, and A. Ephrussi. The Crystal Structure of the *Drosophila* Germline Inducer Oskar Identifies Two Domains with Distinct Vasa Helicase- and RNA-Binding Activities. *Cell Reports*, 12(4):587–598, 2015.
- [22] M. Jeske, C. W. Müller, and A. Ephrussi. The LOTUS domain is a conserved DEAD-box RNA helicase regulator essential for the recruitment of Vasa to the germ plasm and nuage. *Genes Dev.*, 31(9):939–952, May 2017.
- [23] J. Kim-Ha, J. L. Smith, and P. M. Macdonald. *oskar* mRNA is localized to the posterior pole of the *Drosophila* oocyte. *Cell*, 66(1):23–35, July 1991.
- [24] A. Ephrussi and R. Lehmann. Induction of germ cell formation by *oskar*. *Nature*, 358(6385):387–392, July 1992.
- [25] N. Yang, Z. Yu, M. Hu, M. Wang, R. Lehmann, and R.-M. Xu. Structure of *Drosophila* Oskar reveals a novel RNA binding protein. *Proc. Natl. Acad. Sci. U. S. A.*, 112(37):11541–11546, September 2015.

- [26] A. H. Kirk-Spriggs and B. J. Sinclair. *Manual of Afrotropical Diptera Volume 1*. 2017.
- [27] R. S. Peters, L. Krogmann, C. Mayer, A. Donath, S. Gunkel, K. Meusemann, A. Kozlov, L. Podsiadlowski, M. Petersen, R. Lanfear, P. A. Diez, J. Heraty, K. M. Kjer, S. Klopstein, R. Meier, C. Polidori, T. Schmitt, S. Liu, X. Zhou, T. Wappler, J. Rust, B. Misof, and O. Niehuis. Evolutionary History of the Hymenoptera. *Curr. Biol.*, 27(7):1013–1018, April 2017.
- [28] N. F. Vanzo and A. Ephrussi. Oskar anchoring restricts pole plasm formation to the posterior of the *Drosophila* oocyte. *Development*, 129(15):3705–3714, August 2002.
- [29] T. R. Hurd, B. Herrmann, J. Sauerwald, J. Sanny, M. Grosch, and R. Lehmann. Long Oskar Controls Mitochondrial Inheritance in *Drosophila melanogaster*. *Developmental Cell*, 39(5): 560–571, 2016.
- [30] J. A. Gerlt, J. T. Bouvier, D. B. Davidson, H. J. Imker, B. Sadkhin, D. R. Slater, and K. L. Whalen. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta*, 1854(8): 1019–1037, August 2015.
- [31] B. Misof, S. Liu, K. Meusemann, R. S. Peters, A. Donath, C. Mayer, P. B. Frandsen, J. Ware, T. Flouri, R. G. Beutel, O. Niehuis, M. Petersen, F. Izquierdo-Carrasco, T. Wappler, J. Rust, A. J. Aberer, U. Aspöck, H. Aspöck, D. Bartel, A. Blanke, S. Berger, A. Böhm, T. R. Buckley, B. Calcott, J. Chen, F. Friedrich, M. Fukui, M. Fujita, C. Greve, P. Grobe, S. Gu, Y. Huang, L. S. Jermini, A. Y. Kawahara, L. Krogmann, M. Kubiak, R. Lanfear, H. Letsch, Y. Li, Z. Li, J. Li, H. Lu, R. Machida, Y. Mashimo, P. Kapli, D. D. McKenna, G. Meng, Y. Nakagaki, J. L. Navarrete-Heredia, M. Ott, Y. Ou, G. Pass, L. Podsiadlowski, H. Pohl, B. M. von Reumont, K. Schütte, K. Sekiya, S. Shimizu, A. Slipinski, A. Stamatakis, W. Song, X. Su, N. U. Szucsich, M. Tan, X. Tan, M. Tang, J. Tang, G. Timelthaler, S. Tomizuka, M. Trautwein, X. Tong, T. Uchifune, M. G. Walz, B. M. Wiegmann, J. Wilbrandt, B. Wipfler, T. K. F. Wong, Q. Wu, G. Wu, Y. Xie, S. Yang, Q. Yang, D. K. Yeates, K. Yoshizawa, Q. Zhang, R. Zhang, W. Zhang, Y. Zhang, J. Zhao, C. Zhou, L. Zhou, T. Ziesmann, S. Zou, Y. Li, X. Xu, Y. Zhang,

- H. Yang, J. Wang, J. Wang, K. M. Kjer, and X. Zhou. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763–767, November 2014.
- [32] M. Eubanks. Molecular systematics. *Econ. Bot.*, 52(2):133–133, April 1998.
- [33] S. H. Church, J. F. Ryan, and C. W. Dunn. Automation and Evaluation of the SOWH Test with SOWHAT. *Syst. Biol.*, 64(6):1048–1058, November 2015.
- [34] D. Wheeler, A. J. Redding, and J. H. Werren. Characterization of an Ancient Lepidopteran Lateral Gene Transfer. *PLoS ONE*, 8(3):e59262, 2013.
- [35] S. T. Chepkemoi, E. Mararo, H. Butungi, J. Paredes, D. K. Masiga, S. P. Sinkins, and J. K. Herren. Identification of Spiroplasma insolitum symbionts in Anopheles gambiae. *Wellcome Open Research*, 2:90, 2017.
- [36] E. Zchori-Fein, S. J. Perlman, S. E. Kelly, N. Katzir, and M. S. Hunter. Characterization of a ‘Bacteroidetes’ symbiont in Encarsia wasps (Hymenoptera: Aphelinidae): proposal of ‘Candidatus Cardinium hertigii’. *International Journal of Systematic and Evolutionary Microbiology*, 54(3):961–968, 2004.
- [37] E. Zchori-Fein and S. J. Perlman. Distribution of the bacterial symbiont Cardinium in arthropods. *Mol. Ecol.*, 13(7):2009–2016, July 2004.
- [38] K. Bourtzis and T. A. Miller, editors. *Insect Symbiosis, Volume 3 (Contemporary Topics in Entomology)*. CRC Press, 1 edition, October 2008.
- [39] S. Chen, Y. E. Zhang, and M. Long. New Genes in *Drosophila* Quickly Become Essential. *Science*, 330(6011):1682–1685, 2010.
- [40] G. Cornelis, O. Heidmann, S. Bernard-Stoecklin, K. Reynaud, G. Veron, B. Mulot, A. Dupressoir, and T. Heidmann. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. *Proceedings of the National Academy of Sciences*, 109(7):E432–E441, 2012.
- [41] F. Husnik and J. P. McCutcheon. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology*, 16(2):67–79, 2018.

- [42] I. D. Lelio, I. Di Lelio, A. Illiano, F. Astarita, L. Gianfranceschi, D. Horner, P. Varricchio, A. Amoresano, P. Pucci, F. Pennacchio, and S. Caccia. Evolution of an insect immune barrier through horizontal gene transfer mediated by a parasitic wasp. *PLOS Genetics*, 15(3):e1007998, 2019.
- [43] N. Wybouw, Y. Pauchet, D. G. Heckel, and T. Van Leeuwen. Horizontal Gene Transfer Contributes to the Evolution of Arthropod Herbivory. *Genome Biology and Evolution*, 8(6): 1785–1801, 2016.
- [44] D. G. Quispe-Huamanquispe, G. Gheysen, and J. F. Kreuze. Horizontal Gene Transfer Contributes to Plant Evolution: The Case of Agrobacterium T-DNAs. *Frontiers in Plant Science*, 8, 2017.
- [45] L. Boto. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proceedings of the Royal Society B: Biological Sciences*, 281(1777):20132450, 2014.
- [46] A. C. C. Wilson and R. P. Duncan. Signatures of host/symbiont genome coevolution in insect nutritional endosymbioses. *Proceedings of the National Academy of Sciences*, 112(33): 10255–10261, 2015.
- [47] S. López-Madrugal and R. Gil. Et tu, Brute? Not Even Intracellular Mutualistic Symbionts Escape Horizontal Gene Transfer. *Genes*, 8(10):247, 2017.
- [48] N. A. Provorov and O. P. Onishchuk. Microbial Symbionts of Insects: Genetic Organization, Adaptive Role, and Evolution. *Microbiology*, 87(2):151–163, 2018.
- [49] M. Shelomi, E. G. J. Danchin, D. Heckel, B. Wipfler, S. Bradler, X. Zhou, and Y. Pauchet. Horizontal Gene Transfer of Pectinases from Bacteria Preceded the Diversification of Stick and Leaf Insects. *Scientific Reports*, 6(1), 2016.
- [50] Z. Zeng, Y. Fu, D. Guo, Y. Wu, O. E. Ajayi, and Q. Wu. Bacterial endosymbiont *Cardinium* cSfur genome sequence provides insights for understanding the symbiotic relationship in *Sogatella furcifera* host. *BMC Genomics*, 19(1), 2018.
- [51] D. C. Bublitz, G. L. Chadwick, J. S. Magyar, K. M. Sandoz, D. M. Brooks, S. Mesnage, M. S. Ladinsky, A. I. Garber, P. J. Bjorkman, V. J. Orphan, and J. P. McCutcheon. Peptidoglycan

- Production by an Insect-Bacterial Mosaic. *Cell*, 179(3):703–712.e7, 2019.
- [52] X. Xu, J. L. Brechbiel, and E. R. Gavis. Dynein-Dependent Transport of nanos RNA in *Drosophila* Sensory Neurons Requires Rumpelstiltskin and the Germ Plasm Organizer Oskar. *Journal of Neuroscience*, 33(37):14791–14800, 2013.
- [53] A. Jenny. A translation-independent role of *oskar* RNA in early *Drosophila* oogenesis. *Development*, 133(15):2827–2833, 2006.
- [54] J. Rees and K. Cranston. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal*, 5:e12581, 2017.
- [55] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden. NCBI BLAST: a better web interface. *Nucleic Acids Res.*, 36(Web Server issue):W5–9, July 2008.
- [56] P. Grobe. 1KITE - 1K Insect Transcriptome Evolution. March 2017.
- [57] M. A. Benton, N. J. Kenny, K. H. Conrads, S. Roth, and J. A. Lynch. Deep, Staged Transcriptomic Resources for the Novel Coleopteran Models *Atrachya menetriesi* and *Callosobruchus maculatus*. *PLOS ONE*, 11(12):1–23, 12 2016. doi: 10.1371/journal.pone.0167431. URL <https://doi.org/10.1371/journal.pone.0167431>.
- [58] R. C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, August 2004.
- [59] S. Eddy. HMMER. <http://hmmerr.org/>, 2020. Accessed: 2020-10-8.
- [60] M. Stanke, R. Steinkamp, S. Waack, and B. Morgenstern. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32(Web Server):W309–W312, 2004.
- [61] I. Korf. Gene finding in novel genomes. *BMC Bioinformatics*, 5:59, May 2004.
- [62] A. Ahuja and C. G. Extavour. Patterns of molecular evolution of the germ line specification gene *oskar* suggest that a novel domain may contribute to functional divergence in *Drosophila*. *Development Genes and Evolution*, 224(2):65–77, 2014.

- [63] T. U. Consortium and The UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Research*, 37(Database):D169–D174, 2009.
- [64] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O’Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, 31(1):365–370, January 2003.
- [65] P. Shannon. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003.
- [66] S. Capella-Gutierrez, J. M. Silla-Martinez, and T. Gabaldon. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, 2009.
- [67] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.
- [68] J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, August 2001.
- [69] O. Robinson, D. Dylus, and C. Dessimoz. Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. *Molecular Biology and Evolution*, 33(8):2163–2166, 2016.
- [70] A. Löytynoja. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.*, 1079:155–170, 2014.
- [71] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302(1):205–217, September 2000.

Every individual alive today, even the very highest, is to be derived in an unbroken line from the first and lowest forms.

August Weismann

2

Evolutionary history and Functional inference of the Oskar protein

ABSTRACT

Germ line specification is a developmental process essential for the reproduction of sexually reproducing multicellular organisms. In many holometabolous insects, the gene *oskar* is required for the specification of the germ line. However, in hemimetabolous insects, *oskar* plays a role in the nervous system and not in the germ line. To better understand this gene's evolutionary history, and to generate hypotheses addressing how evolutionary changes in protein sequence could have led to changes in the function of Oskar protein, we searched for *oskar* orthologs in 1565 publicly available insect genomic and transcriptomic datasets, and annotated 317 previously undescribed *oskar* orthologs. The earliest-diverging lineage in which we identified an *oskar* ortholog was the order Zygentoma (silverfish and firebrats), suggesting that *oskar* originated before the origin of winged insects (Pterygota). We noted some order-specific trends in *oskar* sequence evolution, including whole gene duplications, clade-specific indels, and rapid divergence. An alignment of all known 379 Oskar sequences revealed highly conserved residues with known biochemical functions in the dimerization of the LOTUS domain. Moreover, we found regions of the OSK domain with conserved predicted RNA binding potential. Furthermore, we show that despite a low overall amino acid conservation, the LOTUS domain shows higher conservation of predicted secondary structure than the OSK domain. Finally, we suggest new key amino acids in the LOTUS domain that may be involved in the previously reported Oskar-Vasa physical interaction.

CONTRIBUTIONS

This work is presented in its draft manuscript form. The work presented here stemmed from the internship of Savandara Besse, a brilliant master student who worked under my, and C. Extavour's supervision for five months. Savandara Besse extended the scripts I wrote in Chapter 1 and created new scripts for the collection and annotation of *oskar* orthologs, she performed the first analyses including the alignments and conservation score computation. She also wrote the first version of the PyMOL visualizer plugin. She computed summary statistics on the genomes and transcriptomes. Finally she wrote the tissue and stage parsing algorithm. Cassandra G. Extavour directed the study design, funding, writing and reviewing. I and C. Extavour proposed the original study designed and I extended and completed the work of Savandara Besse. I repeated the collection of sequences in 2019 and automatized the ortholog detection and filtering, as well as the generation of the alignments. I performed all the phylogenetic analyses presented here. I performed the MCA analysis as well as the secondary structure predictions. I extended and rewrote the scripts created by Savandara Besse for publication purposes. The manuscript and figures were primarily written by me with additions from Savandara Besse. Savandara Besse and Cassandra G. Extavour also extensively reviewed this manuscript.

2.1 Introduction

With the evolution of obligate multicellularity, many organisms faced a challenge considered a major evolutionary transition: allocating only some cells (germ line) to pass on their genetic material to the next generation, relegating the remainder (soma) to death upon death of the organism (reviewed in Kirk¹). This is soma-germline differentiation, where only cells from the germline will create the next generation (reviewed in Kirk¹). While there are multiple mechanisms of germ cell specification, they can be grouped into two broad categories, induction or inheritance². Under induction, cells respond to an external signal by adopting germ cell fate. Under the inheritance mechanism, maternally synthesized cytoplasmic molecules, collectively called germ plasm, are deposited in the oocyte and "inherited" by a subset of cells during early embryonic divisions. Cells inheriting these molecules commit to a germline fate^{2,3}.

The inheritance mechanism in insects that undergo metamorphosis (Holometabola) appears to have evolved by co-option of a key gene, *oskar*⁴. *oskar* was first identified in forward genetic screens for axial patterning mutants in *D. melanogaster*⁴. For the first 20 years following its discovery, *oskar* appeared to be restricted to Drosophilids². Its later discovery in the mosquitoes *Aedes aegypti* and *Anopheles gambiae*⁵ and the wasp *Nasonia vitripennis*⁶ suggested the hypothesis that *oskar* emerged at the base of the Holometabola, and facilitated the evolution of germ plasm in these insects⁶. However, our subsequent identification of *oskar* orthologues in the cricket *Gryllus bimaculatus*⁷, and in many additional hemimetabolous insect species⁸, demonstrated that *oskar* predates the Holometabola, and must be at least as old as the major radiation of insects⁹. Two secondary losses of *oskar* from insect genomes have also been reported, in the beetle *Tribolium castaneum*⁶ and the honeybee *Apis mellifera*¹⁰, and neither of these insects appear to use germ plasm to establish their germ lines^{10,11,12,13}. Whether *oskar* is ubiquitous across all insect orders, whether it is truly unique to insects, the evidence for or against potential losses or duplications of the *oskar* locus across insects, and the evolutionary dynamics of the locus, remain unknown.

oskar remains, to our knowledge, the only gene that has been experimentally demonstrated to

be both necessary and sufficient to induce the formation of functional primordial germ cells^{3,14}. Thus, in *D. melanogaster*^{3,4,14} and potentially more broadly in holometabolous insects with germ plasm^{6,15}, *oskar* plays an essential germ line role. However, it is clear that *oskar*'s germ line function can evolve rapidly, as even within the genus *Drosophila*, *oskar* orthologues from different species cannot always substitute for each other,^{16,17}. Moreover, the ancestral function of this gene may have been in the nervous system rather than the germ line⁷. The current hypothesis is therefore that it was co-opted to play a key role in the acquisition of an inheritance-based germ line specification mechanism approximately 300 million years ago⁹, in the lineage leading to the Holometabola⁷. Thus, the case of *oskar* offers an opportunity to study the evolution of protein function at multiple levels of biological organization, from the genesis of a novel protein, through to potential co-option events and the evolution of functional variation.

Neofunctionalization often correlates with a change in the fitness landscape of the protein sequence caused by novel biochemical constraints imposed by amino acid sequence changes^{18,19}. Such potential constraints may be revealed by analyzing the conservation of amino acids, their chemical properties, or structure at the secondary, tertiary or quaternary levels¹⁹. Oskar has two well-structured domains conserved across identified orthologues to date⁸: an N-terminal Helix Turn Helix (HTH) domain termed LOTUS with potential RNA binding properties^{20,21,22,23}, and a C-terminal GDSL-lipase-like domain called OSK^{20,22,23} (Figure 2.1). These two domains are linked by an unstructured highly variable interdomain sequence^{22,23,24}. We previously showed that this domain structure is likely the result of a horizontal transfer event of a bacterial GDSL-lipase-like domain, followed by the fusion of this domain with a LOTUS domain in the host genome⁸. Biochemical assays of the properties of the LOTUS and OSK domains provide some clues as to the molecular mechanisms that Oskar uses to assemble germ plasm in *D. melanogaster*. The LOTUS domain is capable of homodimerization^{21,22}, and directly binds to and enhances the helicase activity of the ATP-dependent DEAD box helicase Vasa, a germ plasm component²¹. The OSK domain resembles GDSL lipases in sequence^{8,22,23}, but is predicted to lack enzymatic activity, as the conserved amino acid triad (S200 D202 H205) that defines the active site of these lipases is not conserved in OSK^{20,22,23}. Instead,

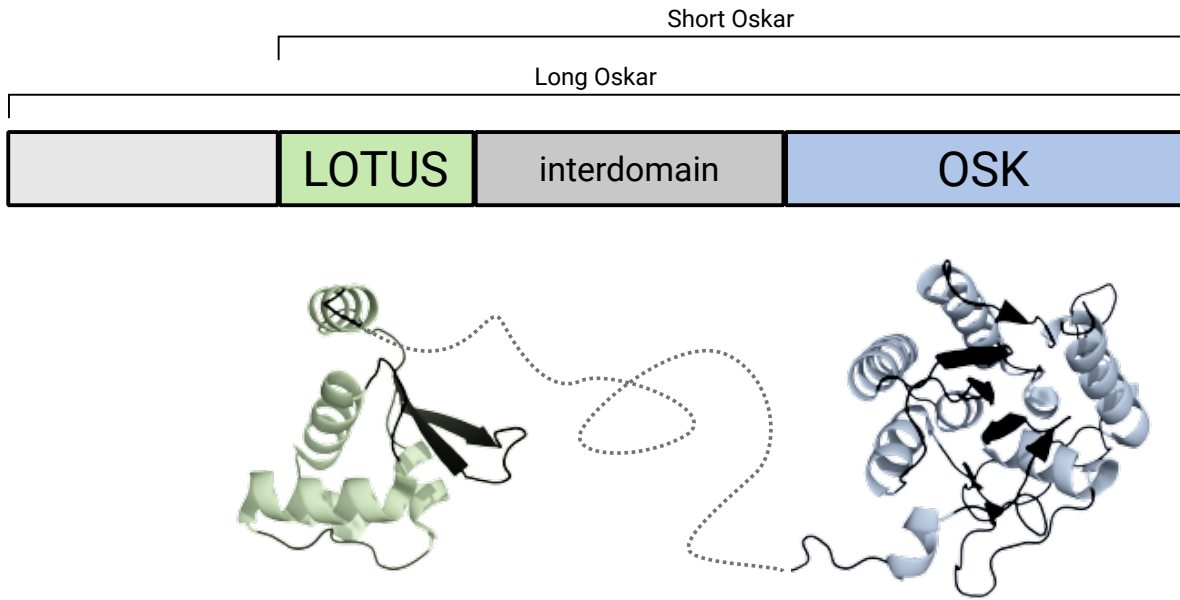


Figure 2.1: Overview of the Oskar protein. Presented in the figure is a schematic representation of the Oskar protein, composed of two folded domains, LOTUS and OSK, separated by an interdomain sequence. Another isoform of the protein is found within some dipteran insects called Long Oskar (as opposed to Short Oskar, the more commonly found isoform). Below the schematic representation is a rendering of the solved structures for LOTUS (PDBID: 5NT7) and OSK (PDBID: 5A4A) with a putative rendering of the unfolded interdomain region.

co-purification experiments suggest that OSK has RNA binding properties, consistent with its predicted basic surface residues^{22,23}. Whether or how changes in the primary sequence of Oskar can explain the evolution of its molecular mechanism or tissue-specific function, remain unknown.

To date, sequences of approximately 100 *oskar* orthologues have been reported^{6,8,22,25}. However, the vast majority of these are from the Holometabola, and it is thus unclear whether analysis of these sequences alone would have sufficient power to allow extrapolation of conservation and divergence of putative biochemical properties across insects broadly speaking. Multiple hypotheses as to the molecular mechanistic function of particular amino acids in the LOTUS and OSK domains in *D. melanogaster* have been proposed^{21,22,23}, but without sufficient taxon sampling, the potential relevance of these mechanisms to *oskar*'s evolution and function in other insects is unclear.

Here we address these outstanding questions by applying a rigorous bioinformatic pipeline to generate the most complete collection of *oskar* sequences to date. By analysing over 1500

Pancrustacean genomes and transcriptomes, we show that *oskar* likely first arose at least 400 million years ago, before the advent of winged insects. We find that the *oskar* locus has been lost independently in some insect orders, including near-total absence from the order Hemiptera, and clarify that the absence of *oskar* from the *Bombyx mori* and *Tribolium castaneum* genomes²⁵ does not reflect a general absence of *oskar* from Lepidoptera or Coleoptera. By comparing Oskar sequences in a phylogenetic context, we reveal that distinct biophysical properties of Oskar are associated with Hemimetabola and Holometabola. We use these observations to propose testable hypotheses regarding the putative biochemical basis of evolutionary change in Oskar function across insects.

2.2 Methods

2.2.1 Experimental model and subject details

This study used no animal model, nor any cell culture lines. However, it used previously generated genomic and transcriptomic datasets. All the information regarding how those datasets were generated can be found on their respective NCBI pages. The list of all the datasets used in this study can be found in the following files: ***genome_insect_database.csv***, ***transcriptome_insect_database.csv***, ***genome_crustacean_database.csv***, and ***transcriptome_crustacean_database.csv***.

2.2.2 Genome and transcriptome preprocessing

We collected all available genome and transcriptome datasets from the NCBI repository registered in September 2019. NCBI maintains two tiers of genomic data: RefSeq, which contains curated and annotated genomes, and GenBank, which contains non-annotated assembled genomic sequences. Transcriptomes are stored in the Transcriptome Shotgun Assembly (TSA) database, with metadata including details on their origin. To search for *oskar* orthologs in datasets retrieved from GenBank, we needed to generate *in silico* gene model predictions. We used the genome annotation tool Augustus²⁶, which requires a Hidden Markov Model (HMM) gene model. To use HMMs producing gene models that would be as accurate as

possible for non-annotated genomes, we selected the closest related species (species with the most recent last common ancestor) that possessed an annotated RefSeq genome. We then used the Augustus training tool to build an HMM gene model for each genome.

We automated this process by creating a series of python scripts that performed the tasks as follows:

- 1) ***1.1_insect_database_builder.py***: This script collects the NCBI metadata regarding genomes and transcriptomes. Using the NCBI Entrez API, it collects the most up to date information on RefSeq, GenBank, and TSA to generate two CSV files: *genome_insect_database.csv* and *transcriptome_insect_database.csv*
- 2) ***1.2_data_downloader.py***: This is a python wrapper around the *rsync* tool that downloads the sequence datasets present in the tables created by (1). It automatically downloads all the available information into a local folder.
- 3) ***1.3_run_augustus_training.py***: This is a python wrapper around the Augustus training tool. It uses the metadata gathered using (1) and the sequence information gathered using (2) to build HMM gene models of all RefSeq datasets. It outputs sbatch scripts that can be run either locally, or on a SLURM-managed cluster. Those scripts will create unique HMM gene models per species.

At the time of this analysis (September 2019), 132 genomes were collected from the RefSeq database, 309 genomes from the GenBank database, and 1114 transcriptomes from the TSA database. All the accession numbers and metadata are available in the two tables (***genome_insect_database.csv*** and ***transcriptome_insect_database.csv***) provided in the supplementary files. This pipeline was repeated for crustaceans (this dataset was downloaded in April 2017) and the information can be found in the following two files: ***genome_crustacean_database.csv*** and ***transcriptome_crustacean_database.csv***.

2.2.3 Creation of protein sequence databases

The classical approach for orthology detection compares protein sequences to amino acid HMM corresponding to the gene of interest. Since we used three different NCBI databases, we

performed the following preprocessing actions:

- 1) RefSeq: well-annotated genomes from NCBI contain gene model translation; no extra processing was required.
- 2) GenBank: Using the HMMs created from the RefSeq databases, we created gene models for each GenBank genome using Augustus and a custom HMM gene model. To choose which HMM gene model to use, we selected the one for each insect order that had the highest training accuracy. In the case where an insect order did not have any member in the RefSeq database, we used the model of the most closely related order. We then translated the inferred coding sequences to create a protein database for each genome. The assignment of the models used to infer the proteins of each GenBank genome is available in the **Table_S2.csv**. To automate the process, we created a custom python script available in the file **1.4_run_augustus.py**.
- 3) TSA: Transcriptomes were translated using the emboss tool Transeq²⁷. We used this tool with the default parameters, except for the six frame translation, trim and clean flags. This generated amino acid sequences for each transcript and each potential reading frame.

2.2.4 Identification of *oskar* orthologs

The *oskar* gene is composed of two conserved domains, LOTUS and OSK, separated by a highly variable interdomain linker sequence^{22,23,24}. To our knowledge, no other gene reported in any domain of life possesses a similar domain composition⁸. Therefore, here we use the same definition of *oskar* orthology as in our previous work, a sequence possessing a LOTUS domain followed by an interdomain region, and then an OSK domain⁸. To maximize the number of potential orthologs, we searched each sequence with the previously generated HMM for the LOTUS and OSK domains⁸. The presence and order of each domain were then verified for each potential hit and only sequences with the previously defined Oskar structure were kept for further processing. We used the HMMER 3.1 tool suite to build the domain HMM (*hmmbuild*

with default parameters), and then searched the generated protein databases (see *Creation of protein sequence databases* above) using those models (*hmmsearch* with default parameters). Hits with an E value ≥ 0.05 were discarded.

All the hits were then aligned with *hmmalign* with default parameters and the HMM of the full-length Oskar alignment previously generated⁸. The resulting sequences were automatically processed to remove assembly artifacts, and potential isoforms. This filtration step was automated and went as follows: First, the sequences were grouped by taxon. Then each group of sequences was aligned using MUSCLE²⁸ with default parameters. The Hamming distance²⁹, a metric that computes the number of different letters between two strings, between each sequence in the alignment was computed. If any group of sequences had a Hamming distance of $>80\%$, then we only kept the sequence with the lowest E-value match. This created a set of sequences containing multiple *oskar* orthologs per species only if they were the likely product of a gene duplication event. We then used the resulting new alignment to generate a new domain HMM and a new full-length Oskar HMM (using *hmmbuild* with default parameters), and ran further iterations of this detection pipeline until we could detect no new *oskar* orthologs in the available sequence datasets. We called this final set the **filtered set** of sequences, and used it in all subsequent orthology analyses unless otherwise specified.

The Oskar sequences obtained are available in the following supplementary files:

Oskar_filtered.aligned.fasta, Oskar_filtered.fasta and Oskar_consensus.hmm.

The domain definitions for the LOTUS and OSK domains are available in the following supplementary files: ***Oskar_filtered.aligned.LOTUS_domain.fasta, LOTUS_consensus.hmm, Oskar_filtered.aligned.OSK_domain.fasta, OSK_consensus.hmm.*** (See *1.5_Oskar_tracker.ipynb*)

2.2.5 Correlative analysis of assembly quality and absence of *oskar*

Using the metadata gathered previously (see *Genomes and transcriptomes preprocessing* above) we created two pools of source data: genomes where we found an *oskar* sequence, and genomes where we failed to find a sequence that met our orthology criteria. We then compared the two distributions for each of the 8 available assembly statistics: (1) Contig and (2) Scaffold N50, (3)

Contig and (4) Scaffold L50, (5) Contig and (6) Scaffold counts, and (7) Number of Contigs and (8) Scaffolds per genome length. Finally, we performed a Mann-Whitney U statistical analysis to compare the means of the two distributions (See **2.1_Oskar_discovery_quality.ipynb**).

2.2.6 TSA metadata parsing and curation

Datasets in the TSA database are associated with a biosample object that contains all the metadata surrounding the RNA sequencing acquisitions. These metadata can include information about one or both the tissue of origin and the organism's developmental stage. We first automated the retrieval of these metadata using a custom python script that used the NCBI Entrez API (see **2.3_Oskar_tissues_stages.ipynb**). However, the metadata proved to be complex to parse: (1) not all projects had the data entered in the corresponding tag, (2) some data contained typographical errors, and (3) multiple synonyms were used to describe the same thing with different words in different datasets. We therefore created a custom parsing and cleaning pipeline that corrected mistakes, and aggregated them into a cohesive set of unique terms that we thought would be most informative to interpret the presence or absence of *oskar* orthologs (see **2.3_Oskar_tissues_stages.ipynb** to see the mapping table). This strategy sacrificed some of the fine-grained information contained in custom metadata (for example "right leg" became "leg"), but allowed us to analyze the expression of *oskar* uniformly throughout all the datasets. This pipeline generated, for all available datasets, a table of tissues and developmental stages containing *oskar* transcripts (see **Oskar_all_tissues_stages.csv**).

2.2.7 Dimensionality reduction of Oskar alignment sequence space

The Oskar alignment was subjected to a Multiple Correspondence Analysis (MCA). Similar to a PCA, dimension vectors are first computed to maximize the spread of the underlying data in the new dimensions, except that instead of a continuous dataset, each variable (here an amino acid at a given position) contributes to the continuous value on that dimension. Once the projection vectors are computed, each sequence is then mapped onto the dimensions. Each amino acid position (column) in the alignment was considered a dimension with a possible value set of 21 (20 amino acids and gap). We first removed the columns of low information (columns that had

less than 30% amino acid occupancy) using *trimal*³⁰ with a cutoff parameter set at 0.3. Then, the alignment was then decomposed into its eigenvectors, and projected to the first 3 components. To perform this decomposition, we implemented a previously developed preprocessing method³¹ in a python script (see *MCA.py* and *2.8_Oskar_MCA_Analysis.ipynb*) and performed the eigenvector decomposition with the previously developed MCA python library (see *Key resource table*). We ran the same algorithm on the LOTUS domain, OSK domain, and full-length Oskar alignments obtained above (see *Identification of oskar orthologs above*).

2.2.8 Phylogenetic inference of Oskar sequences in the Hymenoptera

We aligned all Hymenopteran Oskar sequences using PRANK³² with default parameters. We then used the result of the filtering steps presented in *Identification of oskar orthologs* to detect duplicated *oskar* sequences. If two or more *oskar* sequence were present within the same species we annotated them as duplicated. We trimmed this alignment to remove all columns with less than 50% occupancy using *trimal* with the cutoff parameter set at 0.5. To reconstruct the phylogeny of these sequences, we used the maximum likelihood inference software RAxML³³ with a gamma-distributed protein model, and activated the flag for auto model selection. We ran 100 bootstraps and then visualized and annotated the obtained tree with Ete3³⁴ in a custom ipython notebook (see *2.7_Oskar_duplication.ipynb*).

2.2.9 Calculation of Oskar conservations scores

Using the large set of orthologous Oskar sequences we obtained as described above, we computed different conservation scores for each amino acid position. This methodology relies on the hypotheses that if an amino acid, or the chemical properties associated with it, at a particular position in the sequence are important for the structure and / or function of the protein, they will be conserved across evolution. We considered multiple conservation metrics, each highlighting a particular aspect of the protein properties : (1) Mixed biochemical property conservation scores with the Valdar³⁵ and Jensen-Shannon Divergence (JSD)^{36,37} scores (2) Electrostatic charge conservation (3) Hydrophobicity conservation (4) RNA Binding prediction conservation (5) Secondary structure prediction conservation. While we report the JSD score in

the table, we did not use it for the analysis presented in this study. The scores can be found in the supplementary file: **scores.csv**.

Computation of the Valdar score

The Valdar score attempts to account for transition probabilities, stereochemical properties, amino acid frequency gaps, and, essential for this study, sequence weighting. Due to the heterogeneity of sequence dataset availability, most Oskar sequences occupy only a small portion of insect diversity, primarily Hymenoptera and Diptera. Sequence weighting allows for the normalization of the influence of each sequence on the score based on how many similar sequences are present in the alignment³⁵. We implemented the algorithm described in Valdar³⁵ in a python script (see **besse_blonde_l_conservation_scores.py**), then calculated the conservation scores for the Oskar alignment we generated above.

Computation of the Jensen-Shannon Divergence score

Jensen-Shannon Divergence (JSD) uses the amino acid properties and stereochemical properties together to infer the "amount" of evolutionary pressure an amino acid position is subject to. This score uses an information theory approach by measuring how much information (in bits) any position in the alignment brings to the overall alignment³⁶. This score also takes into account neighboring amino acids in calculating the importance of each amino acid. We used previously published python code to calculate the JSD of our previously generated Oskar alignment³⁶ (see **score_conservation.py**).

Computation of the Conservation Bias

The measure of differences in conservation between the holometabolous and hemimetabolous Oskar sequences presented in the results was done as follows: we first split the alignment into two groups containing the sequences from each clade (see **2.4_Oskar_pgc_specification.ipynb**). Due to the high heterogeneity in taxon sampling between hemimetabolous and holometabolous insects, we ran a bootstrapped approximation of the conservation scores on holometabolous sequences. We randomly selected N sequences (N = the number of hemimetabolous sequences), computed the Valdar conservation score (see *Computation of the Valdar score above*), and stored it.

After 1000 iterations, we computed the mean conservation score for each position for holometabolous sequences. For hemimetabolous sequences, we directly calculated the Valdar score using the method as described above (see *Computation of the Valdar score*). For each position, we then computed what we refer to as the "conservation bias" between Holometabola and Hemimetabola by taking the ratio of the log of the conservation score Holometabola and Hemimetabola. $ConservationBias = \frac{\log Valdar_{holo}}{\log Valdar_{hemi}}$ for each position. (see

3.4_LogRatio_Bootstrap.ipynb)

Computation of the electrostatic conservation score

To study the conservation of electrostatic properties of the Oskar protein we computed our own implementation of an electrostatic conservation score (see ***besse_blonde_l_conservation_scores.py***). Aspartic acid and Glutamic acid were given a score of -1, Arginine and Lysine a score of 1, and Histidine a score of 0.5. All other amino acids were given a score of 0. Then, we summed the electrostatic score for each sequence at each position, and divided this raw score by the total number of sequences in the alignment. This computation assigns a score between -1 and 1 at each position, -1 being a negative charge conserved across all sequences, and 1 a positive charge.

Computation of the hydrophobic conservation score

To study the conservation of hydrophobic properties of the Oskar protein we implemented our own hydrophobic conservation score (see ***besse_blonde_l_conservation_scores.py***). At each position, each amino acid was given a hydrophobic score taken from a previously published scoring table³⁸ (This table is implemented in the ***besse_blonde_l_conservation_score.py*** file for simplicity). Scores at each position were then averaged across all sequences. This metric allowed us to measure the hydrophobicity conservation of each position in the alignment, and is bounded between 5.39 and -2.20.

Computation of the RNA binding affinity score

RNA binding sites are defined as areas with positively charged residues and hydrophobic residues. To estimate the conservation of RNA binding sites in *oskar* orthologs, we used

RNABindR³⁹, an algorithm predicting putative RNA binding sites based on sequence information only. We automated the calculation for each sequence by writing a python script that submitted a request to the RNABindR web service (see *RNABindR_run_predictions.py*). We then aggregated all results into a scoring matrix, and averaged the score obtained for each position. We call this score the RNABindR score, and hypothesize that it reflects the conservation of RNA binding properties of the protein. Importantly, this score was obtained in 2017 for only a subset of 219 proteins used in this study. Since then, the RNABindR server has been defunct and we could not repeat those measurements as the source code for this software is unavailable.

Computation of secondary structure conservation

Due to the overall low conservation of the LOTUS domain, we decided to see whether the secondary structure was conserved. To this end, we used the secondary structure prediction algorithm JPred 4⁴⁰, a tool which, given an amino acid sequence, returns a positional prediction for α -helix, β -sheet or unstructured. We used the JPred4 web servers to compute the predictions and processed them into a secondary structure alignment (see *2.6_Oskar_lotus_osk_structures.ipynb*). We then used WebLogo⁴¹ to visualize the conservation of the secondary structure.

2.2.10 Visualization of conservation scores

We used Pymol⁴² to look at the different computed conservation scores mapped onto the solved structures of LOTUS and OSK^{21,22}. At the time of writing, no full-length Oskar protein structure had been reported. With the caveat that all visualization was done on the structure of the *D. melanogaster* protein, we created a custom python script that augments pymol with automatic display and coloring capacities. This script is available as *Oskar_pymol_visualization.py*, and contains a manual at the beginning of the file. For the OSK domain, we used the structure PDBID: 5A4A, and for the LOTUS domain, PDBID: 5NT7^{21,22}. The LOTUS structure we used is in complex with Vasa, and in a dimeric form²¹, allowing for easy interpretation of the different conservation scores. For the OSK structure, we removed the residues 399-401 and 604-606 from

the PDB file as those amino acids did not align across all sequences and therefore showed highly biased conservation scores.

2.2.11 Quantification and statistical analysis

Statistical analysis

All statistical analyses were performed using the scipy stats module (<https://www.scipy.org/>). Significance thresholds for p-values were set at 0.05. Statistical tests and p-values are reported in the figure legends. All statistical tests can be found in the ipython notebooks mentioned below.

2.2.12 Data and code availability

The study generated a series of python3 script and python 3 ipython notebook files that perform the entire analysis. All the results presented in this paper can be reproduced by running the aforementioned python3 code. The primary data, *oskar* orthologs, Oskar alignments, trees, and conservation statistics as well as the code created and used are available as supplementary information. For ease of access, legibility, and reproducibility, the code and datasets have been deposited in a GitHub repository available at https://github.com/extavourlab/Oskar_Orthologs

2.2.13 Key resources table

Software and libraries

All software and libraries used in this study are published under open source libre licenses and are therefore available to any researcher.

Type	Name	Version	Source
Software	HMMER	3.1.b2	http://hmmer.org/
Software	Pymol	1.8.x	https://pymol.org
Software	rsync	3.1.2	http://rsync.samba.org/
Software	Python3	3.7	https://www.python.org/
Software	MrBayes	3.2.6	http://nbisweden.github.io/MrBayes/

Type	Name	Version	Source
Software	trimal	1.2rev59	http://trimal.cgenomics.org/
Software	transeq	6.6.0.0	http://emboss.sourceforge.net/
Software	augustus	2.5.5	http://augustus.gobics.de/
Software	JPred4	4.0	http://www.compbio.dundee.ac.uk/jpred/
Software	RNABindR	2.0	http://ailab1.ist.psu.edu/RNABindR/
Software	Inkscape	0.92.3	https://inkscape.org/
Library	jupyter	4.4.0	https://jupyter.org/
Library	ete3	3.3.1	http://etetoolkit.org
Library	pandas	0.25.1	https://pandas.pydata.org/
Library	mca	1.0.3	https://pypi.org/project/mca/
Library	fuzzywuzzy	0.17.0	https://github.com/seatgeek/fuzzywuzzy
Library	BeautifulSoup4	4.6.3	https://pypi.org/project/beautifulsoup4/
Library	biopython	1.74	https://pypi.org/project/biopython/
Library	numpy	1.16.2	https://www.numpy.org/
Library	seaborn	0.9.0	https://seaborn.pydata.org/
Library	matplotlib	3.0.0	https://matplotlib.org/
Library	scipy	1.1.0	https://www.scipy.org/
Library	progressbar	3.38.0	https://github.com/niltonvolpato/python-progressbar

2.3 Results

2.3.1 HMM-based discovery pipeline yields hundreds of novel *oskar* orthologs

We wished to study the evolution of the *oskar* gene sequence as comprehensively as possible across all insects. To expand our previous collection of nearly 100 orthologous sequences⁸, we designed a new bioinformatics pipeline to scan and search for *oskar* orthologs across all 1565 NCBI insect transcriptomes and genomes that were publicly available at the time of analysis

(Figure 2.2; see Methods: *Genome and transcriptome pre-processing* for NCBI accession numbers and additional information). First, we used the HMMER tool suite to build HMM models for each of the LOTUS and OSK domains, using our previously generated multiple sequence alignments (MSA)⁸. We subjected genomes to *in silico* gene model inference using Augustus²⁶. We translated the resulting predicted transcripts, as well as the predicted transcripts from RNA-seq datasets, in all six frames. We then scanned the resulting protein sequences for the presence of LOTUS and OSK domains using the aforementioned HMM models. Sequences were designated as *oskar* orthologs based on the same criteria as in our previous study⁸, namely, sequences containing both a LOTUS and OSK domain²², separated by a variable interdomain region. We then aligned all sequences using *hmmalign* and the HMM derived from our previously published full length Oskar alignment⁸, and manually curated sequence duplicates and sequences that did not align correctly.

With these methods, we recovered a total of 379 unique *oskar* sequences from 350 unique species, 317 of which were previously unannotated. To our knowledge, this comprises the largest collection of *oskar* orthologs described to date. To determine if *oskar* orthologues might predate Insecta, we applied the discovery pipeline to all 31 non-insect pancrustacean genomes and 266 transcriptomes available at the time of analysis (see methods: *Genomes and transcriptomes preprocessing* for complete list). However, we did not recover any non-insect sequences meeting our criteria for *oskar* orthologs (Figure 2.3), strongly suggesting that *oskar* is restricted to the insect lineage^{6,24}.

We found that 58% of RefSeq genomes, 31% of GenBank genomes, and 19% of transcriptomes analysed contained predicted *oskar* orthologs (Table B.1 and Figure B.1a). Given that detection of putative orthologs is highly dependent on the quality of the genome assembly and annotation, we asked whether there were differences in the assembly statistics of genomes with and without predicted *oskar* orthologs. We observed a significant difference in N50, L50, number of contigs and number of scaffolds between genomes lacking *oskar* hits and those where *oskar* was found (Mann-Whitney U test p-value < 0.05). Genomes where we did not find *oskar* showed a higher contig (255015 vs 43280) and scaffold count (182706 vs 23596), a smaller N50 for contigs (324036bp vs 726696bp) and scaffolds (2636825bp vs 5695299bp), a larger L50 for

contigs (40955 vs 3701) and scaffolds (27269 vs 1500), and a larger number of contigs (0.00060 vs 0.00017) or scaffolds (0.00045 vs 0.00009) per genome length, than genomes where we detected an *oskar* ortholog (as shown in Figure B.2).

Figure 2.2 (following page): Presentation of the *oskar* ortholog detection pipeline. Sequences were collected automatically from the three NCBI databases, Genbank (GCA), RefSeq (GCF) and Transcriptome Shotgun Assemblies (TSA). RefSeq genomes were used to generate Augustus gene model HMM which were used to annotate and predict proteins in the GenBank non annotated genomes. Transcripts from the TSA database were 6-frame translated using TRANSEQ. Amino acid sequences were consolidated into three protein databases. *hmmsearch* from the HMMER tool suite was used to search for LOTUS and OSK hits in those sequences. Sequences with both a LOTUS and OSK hit with a E-value < 0.05 were kept and annotated as Oksar sequences. Sequences were then cleaned to remove duplicates (sequences with > 80% sequence similarity coming from the same organism). Remaining sequences were then aligned using *hmmalign*, and the process was repeated until no new sequences were found. Finally, the sequences were consolidated with the dataset metadata into the *oskar* ortholog database that is used for the rest of the analysis.

Figure 2.2: (continued)

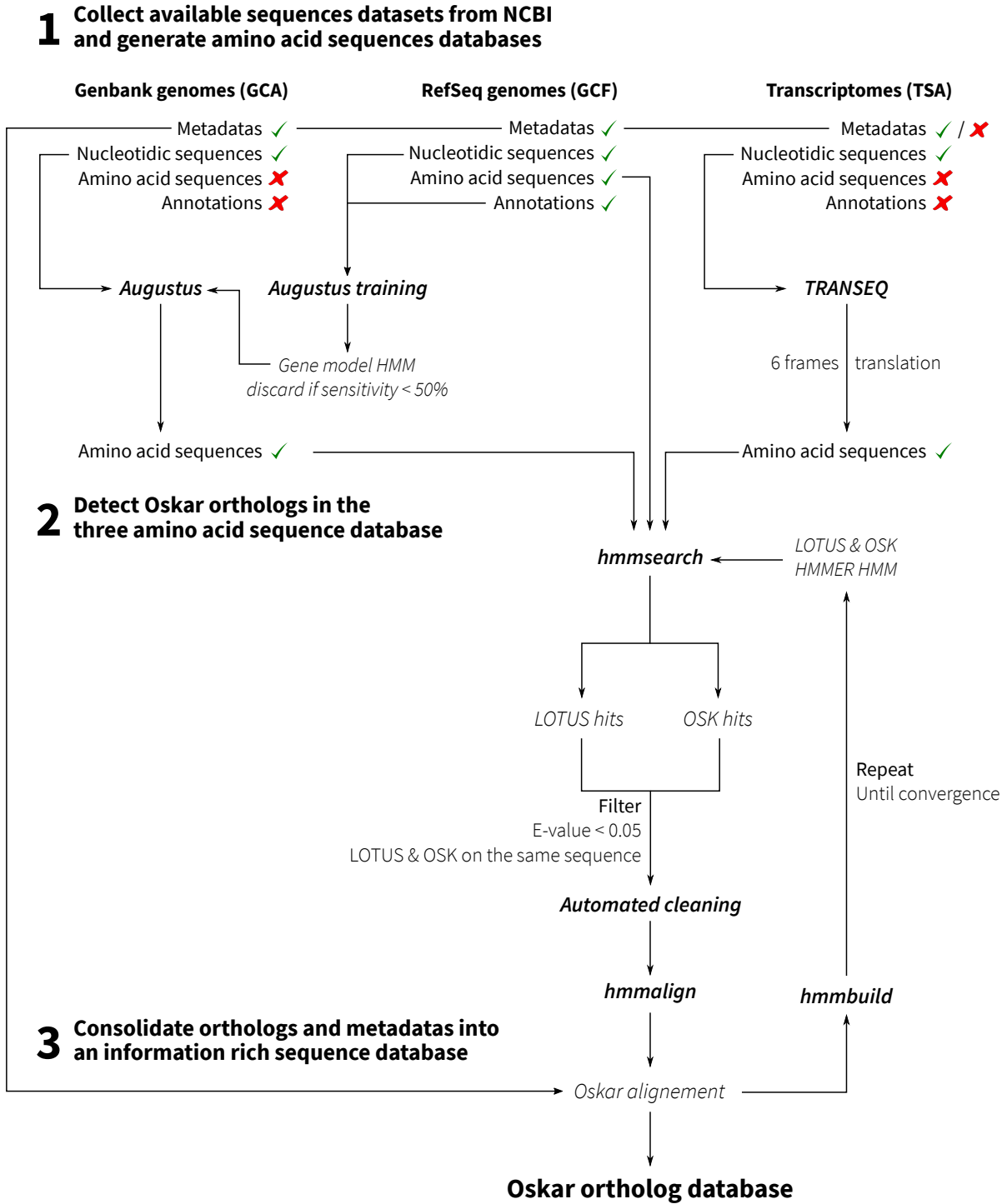
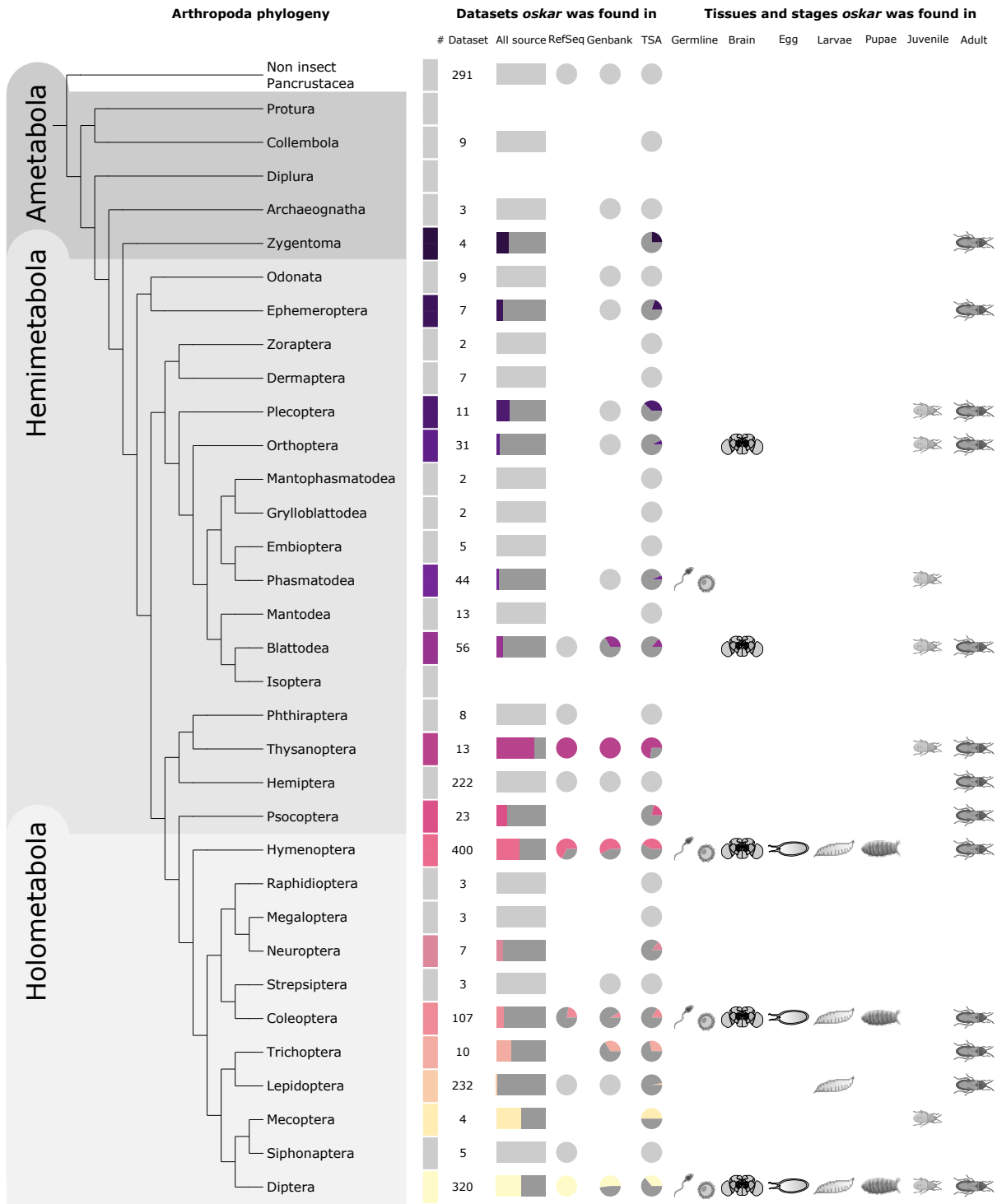


Figure 2.3 (following page): Summary of *oskar* distribution and expression in insects. Phylogeny from Misof et al. ⁹. In order from left to right: Color code for the insect order, if grey, no *oskar* was found in this order. Number of datasets searched. Proportion and absolute number of *oskar* sequences found. Proportion and absolute number of *oskar* sequences found in RefSeq datasets. Proportion and absolute number of *oskar* sequences found in Genbank datasets. Proportion and absolute number of *oskar* sequences found in TSA datasets. *oskar* sequences found in tissue related to germline (reproductive organs + eggs). *oskar* sequences found in tissue related to the brain (neuronal, brain and head). *oskar* sequences found in an egg transcriptome. *oskar* sequences found in a larval transcriptome. *oskar* sequences found in a pupal transcriptome. *oskar* sequences found in a nymph/juvenile transcriptome. *oskar* sequences found in an adult transcriptome. All numbers represented here can be found in the Table B.1.

Figure 2.3: (continued)



2.3.2 *oskar* predates the divergence of Ametabola and other insects

We found *oskar* orthologs in 14 of the 29 generally recognized⁹ insect orders, including six holometabolous orders, seven hemimetabolous orders, and one ametabolous order. This result is consistent with our previous finding that *oskar* predates the origins of the Holometabola^{7,8,24}. The novel finding of an *oskar* ortholog from the silverfish *Atelura formicaria* (Zygometa) allows us to date back the origin of *oskar* further than previous analyses, to at least 420 million years ago⁹, before the divergence of Ametabola from the remaining insect lineages.

We then explored the distribution of *oskar* sequences across insect phylogeny. Interestingly, we found multiple lineages where *oskar* appeared to have been lost independently, including confirming the previously reported⁶ losses from the genomes of the red flour beetle *Tribolium castaneum*, the honeybee *Apis mellifera*, and the silk moth *Bombyx mori* (Figure 2.3). Notably, within Lepidoptera we found *oskar* orthologues in only four species, despite the fact that we searched 232 available lepidopteran sequence datasets (Figure 2.3 and Figure B.3), including 17 well-annotated RefSeq genomes and 135 transcriptomes. In principle, this apparent widespread absence of *oskar* in Lepidoptera could be due to unusually rapid evolution of the *oskar* sequence in this lineage, which might render lepidopteran *oskar* orthologues undetectable by our methods. However, we note that the only four lepidopteran orthologs we detected all belonged to species of the basally branching *Adelidae* and *Palaephatidae* families. We therefore favor the interpretation that *oskar* was lost from a last common ancestor of *Meessiidae* and *Palaphaetidae*, approximately 180 million years ago, with the consequence that the majority of extant lepidopteran lineages lack an *oskar* ortholog (Figure B.3)^{43,44}.

The Hemiptera also appear to have lost *oskar*, based on our analysis of the 222 datasets available for this clade, including 12 RefSeq genomes and 192 transcriptomes. However, we did identify an *oskar* ortholog in the Thysanoptera, which is a hemipteran sister group⁹. Finally, we found *oskar* orthologs in only four of the 11 orders of the Polyneoptera for which data were available. With the exception of Mantodea (13 transcriptomes), the four orders with detectable *oskar* sequences all had more than 10 available sequence datasets (Plecoptera: 3 genomes and 8 transcriptomes; Orthoptera: 3 genomes and 28 transcriptomes; Phasmatodea: 13 genomes and

31 transcriptomes; Blattodea: 5 genomes and 51 transcriptomes). The remaining six orders had fewer than eight datasets each available for analysis (Figure 2.3; Table B.1), which could account for the apparent paucity of *oskar* genes in this group. However, we cannot rule out the possibility that *oskar* in the polyneoptera may have diverged beyond our ability to detect it, or that it may have been lost multiple times, as observed for multiple holometabolous orders.

As well as multiple convergent losses of *oskar*, we also uncovered evidence for independent instances of duplication of the *oskar* locus. We defined a putative duplication instance as two or more *oskar* sequences (possessing both a LOTUS and OSK domain as per our definition) in the same species that shared less than 80% sequence similarity. All of these events were detected within the Hymenoptera. We therefore performed a phylogenetic analysis of the hymenopteran sequences to test the hypothesis that these were the result of duplication events (Figure 2.4; Figure B.4). Our analysis recovered previously published hymenopteran phylogenetic relationships⁴⁵. We found that *oskar* was duplicated in the four *Figitidae* species studied, a family of parasitoid wasps. Moreover, one out of two *Cynipidae* species, as well as the only *Ceraphronidae* species examined, also harbored a duplicated *oskar* sequence. This suggests that a duplication of *oskar* occurred in the common ancestor of those three families. Multiple *oskar* duplications were also found in the Chalcid wasps (Chalcidoidea families), notably in the *Mymaridae* (three out of four species), the *Eupelmidae* (two out of three species), the *Aphelinidae* (two out of two species) and the *Pteromalidae* (one out of twelve species). We suggest that a duplication of *oskar* in a common ancestor of the Chalcid wasps is more parsimonious than multiple duplication events. Finally, two isolated duplication events were found in the *Aculeata*, one in the *Vespidae* family in the wasp *Polistes fuscatus*, and one in the *Formicidae* in the red imported fire ant *Solenopsis invicta*.

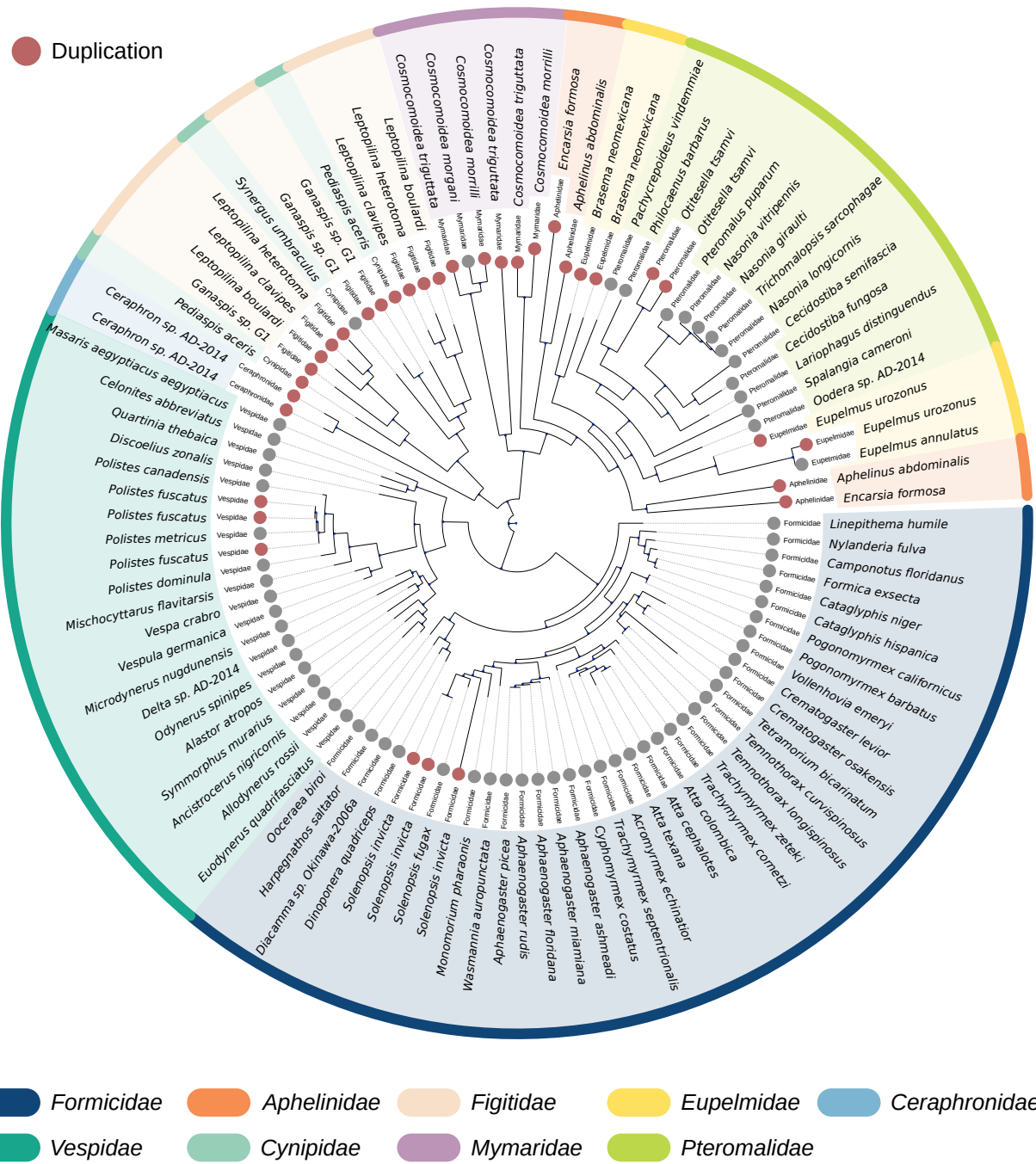


Figure 2.4: Phylogenetic reconstruction of hymenopteran Oskar sequences. Phylogenetic tree inferred using RaxML with 100 bootstrap. Each leaf is an Oskar ortholog. In gray, only one Oskar sequence was found in this species, in red duplicated Oskars sequences (sequence similarity < 80%). Only the families which displayed a putative duplication are shown here, see Figure B.4 for complete hymenopteran phylogeny.

2.3.3 Evidence for *oskar* expression in multiple somatic tissues

In studied insects to date (the cricket *G. bimaculatus*⁷ and the fruit fly *D. melanogaster*⁴⁶), *oskar* is expressed and required in one or both of the germ line^{3,5,6} or the nervous system. We asked whether these expression patterns could be generalized across the insects studied here. To this end, we downloaded all available metadata for the transcriptomes analysed here, to obtain information on the source tissues and developmental stages. We obtained these data for 371 out of the 1164 transcriptomes in our analysis, including both holometabolous and hemimetabolous orders. To first explore the distribution of *oskar* expression in the brain and the germline, we binned the different tissues reported in the metadata into two categories, brain or germline. This was done independently of the developmental stage (if that information was included in the metadata) by creating a mapping table and checking the extracted tissues against this table (see Methods: *TSA metadata parsing and curation*). We then cross referenced our orthology detection with these metadata. We found evidence for *oskar* expression in the germ line of four orders (Phasmatodea, Hymenoptera, Coleoptera and Diptera), and in the brain of five orders (Orthoptera, Blattodea, Hymenoptera, Coleoptera, Diptera) (see Methods: *TSA metadata parsing and curation* for details on keyword extractions). In addition, we found evidence of *oskar* expression in a number of somatic tissues not previously implicated in studies of *oskar* expression and function. These tissues included the midgut (*Polistes fuscatus*, *Sitophilus oryzae*), fat body (*Polistes fuscatus*, *Arachnocampa luminosa*), salivary gland (*Culex tarsalis*, *Anopheles aquasalis*, *Leptinotarsa decemlineata*), venom gland (*Culicoides sonorensis*, *Fopius arisanus*), and silk gland (*Bactrocera cucurbitae*) (Figure B.5). In terms of developmental stage, only holometabolous insects appeared to express *oskar* during embryonic, larval or nymphal stages; for all other insects, *oskar* was detected in transcriptomes derived from adults (Figure 2.3). However, it is important to note that for most species, transcriptomes are available only from adult tissues, rather than from a full range of developmental stages (Figure B.5). We therefore cannot rule out the possibility that *oskar* expression at pre-adult stages is also a feature of multiple Hemimetabola. Indeed, we previously reported that *oskar* is expressed and required in the embryonic nervous system of a cricket, a hemimetabolous insect⁷.

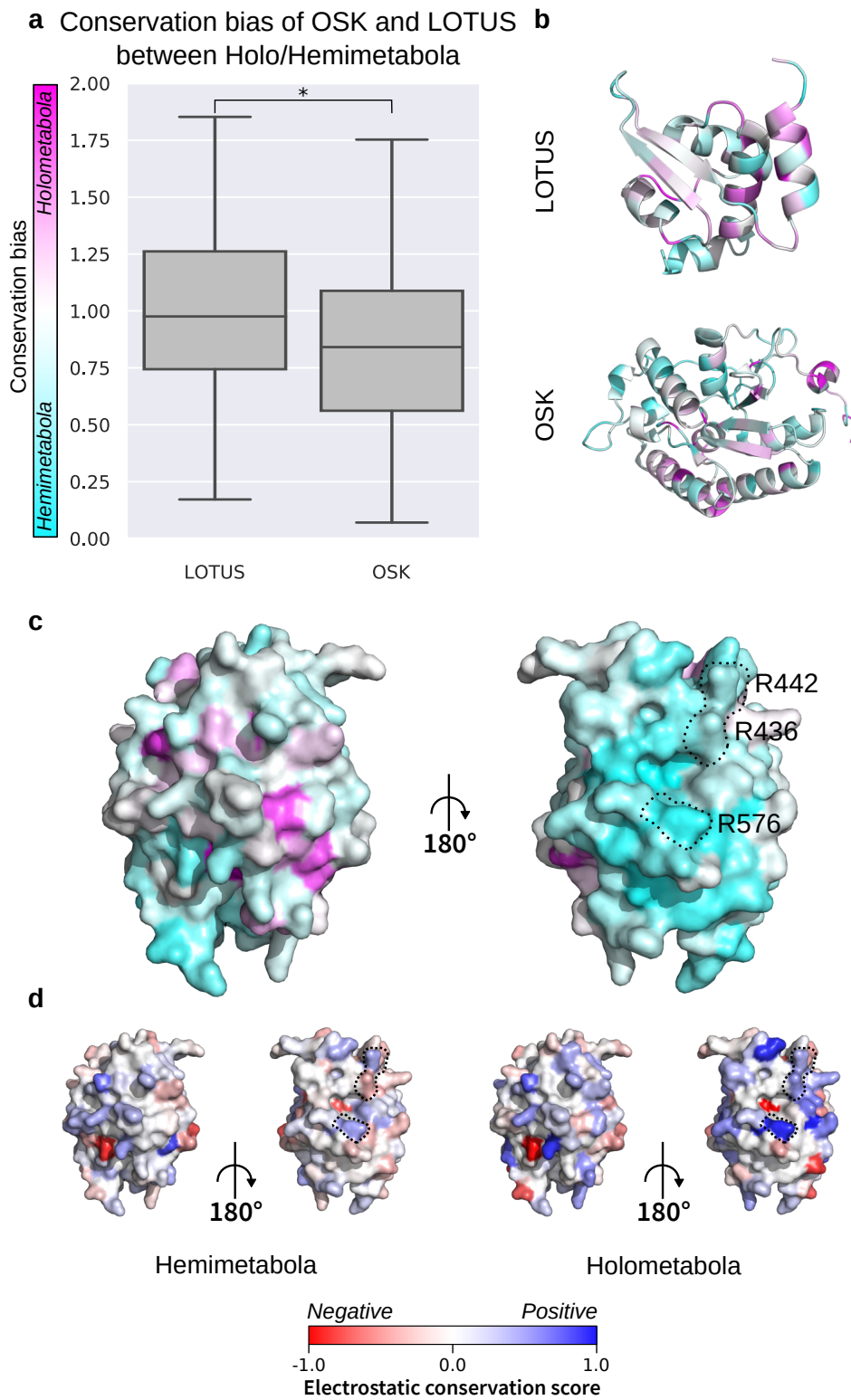
2.3.4 LOTUS and OSK evolved differently between hemimetabolous and holometabolous insects

The fact that an *oskar*-dependent germ plasm mode of germ line specification mechanism is found only in holometabolous insects suggests that *oskar* may have been co-opted in this clade for this function (discussed in Ewen-Campen et al. ⁴⁷). Under this hypothesis, evolution of the *oskar* sequence in the lineage leading to the Holometabola may have changed the physico-chemical properties of Oskar protein, such that it acquired germ plasm nucleation abilities in these insects. To test this hypothesis, we asked whether there were particular sequence features associated with Oskar proteins from holometabolous insects, in which Oskar can assemble germ plasm, and hemimetabolous insects, which lack germ plasm. In particular, we assessed the differential conservation of amino acids at particular positions across Oskar, and asked if these might be predicted to change the physico-chemical properties of Oskar in specific ways that could potentially be relevant to germ plasm nucleation. We decided to use the Valdar score ³⁵ as the main conservation indicator for this study (Supplementary File *scores.csv*). The Valdar score accounts not only for transition probabilities, stereochemical properties and amino acid frequency gaps, but also for the availability of sequence diversity in the dataset. It computes a weighted score, where less represented sequences participate to a greater effect to the score than overrepresented sequences. Due to the highly unbalanced sampling between hemimetabolous and holometabolous sequences the choice of a weighted score was necessary to avoid biasing the results towards insect orders such as Diptera or Hymenoptera. In order to study the difference between hemimetabolous and holometabolous sequences, we did not use the Valdar score directly, but instead computed the conservation ratio between both groups for each position, which we call the Conservation bias (See Methods: Computation of the Conservation Bias). We plotted the conservation bias on the solved three-dimensional crystal structure of the *D. melanogaster* LOTUS and OSK domain ^{22,23} to ask whether specific functionally relevant structures showed phylogenetic or other patterns of residue conservation (Figure 2.5). First, we asked if the overall conservation score of the domains was different between holometabolous and hemimetabolous sequences. We observed that the conservation bias for

the LOTUS domain was centered around a mean of 1.00, indicating that both holometabolous and hemimetabolous displayed a similar conservation of the LOTUS domain (Figure 2.5a). For the OSK domain however, the conservation bias was centered around 0.84, indicating that the hemimetabolous sequences displayed a higher level of conservation compared to holometabolous sequences (Figure 2.5a). We then looked at the conservation bias scores *in-situ* on the LOTUS domain structure. We asked if the amino acids of the β sheets of the LOTUS domain thought to be involved in dimerization of the protein²² displayed conservation bias. Both β sheets had an overall even bias (mean: 1.03 and 1.05 for β 1 and β 2 respectively) between both groups (Figure 2.5b). Second, as we had observed that the OSK domain had an overall biased conservation, with hemimetabolous OSK showing a higher conservation overall, we asked if there were any clear pattern of conservation bias in the structure (Figure 2.5a and b). Some of the secondary structures showed a differential conservation (α 2: 0.54, α 6: 0.42, β 2: 0.52), whereas the other structures were within less than 0.1 of the median value for OSK. Moreover, we observed a large pocket of amino acids showing a conservation bias towards hemimetabolous sequences located on the surface of OSK (Figure 2.5c). This particular area contains the previously reported important amino acids for the RNA binding function of OSK^{22,23} namely, R442, R436 and R576. Surprisingly, when we looked at the differences in electrostatic conservation between holometabolous and hemimetabolous sequences, we noticed that the electrostatic properties at those positions were conserved in the holometabolous sequences R436:0.36, R442:0.29 and R576:0.81 (Figure 2.5d).

Figure 2.5 (following page): Differential conservation of amino acids between hemimetabolous and holometabolous sequences. In **a**), boxplot showing the conservation bias between hemimetabolous and holometabolous sequences. Represented is the conservation bias for each of the two domains of Oskar. Statistical difference tested using MannWhitney U test ($p < 0.05$). In **b**), cartoon representation of LOTUS (PDBID: 5NT7) and OSK (PDBID: 5A4A) where each amino acid is colored by its conservation bias. In **c and d**), protein surface representation of the OSK (PDBID: 5A4A) domain. Circled with black dashed lines are the three amino acids reported previously to be necessary for OSK binding to RNA in *D. melanogaster*^{22,23}. In **c**), amino acids are colored by their conservation bias. In cyan, amino acids show higher conservation in hemimetabolous sequences while in purple they are more conserved in holometabolous sequences. In **d**), amino acids are colored by their electrostatic conservation score. On the left, hemimetabolous sequences and on the right holometabolous sequences.

Figure 2.5: (continued)



To gain further insight into the differences in conservation across insects, we reduced the multiple sequence alignment dimensionality using a Multiple Correspondence Analysis (MCA), an equivalent of PCA for categorical variables⁴⁸. We performed the dimensionality reduction for the full length Oskar sequence alignment as well as for the LOTUS and OSK alignments (Figure B.6). Interestingly, we found that most of the variance in sequence space was due to dipterans and hymenopterans (Figure B.6). When we considered the OSK domain only, we found clusters of *Drosophilidae*, *Culicidae* and *Formicidae* sequences (Figure B.6). This clustering is also reflected for the LOTUS domain, where the *Drosophilidae* and *Culicidae* contribute to a high amount of variance in the first MCA dimension. However, for the LOTUS domain, the *Formicidae* sequences do not cluster away from other Oskar sequences (Figure B.6). This suggests that the LOTUS domain of Diptera diverged in sequence between *Drosophilidae* and *Culicidae*.

Finally, we examined the origins and evolutionary dynamics of Oskar isoforms. *D. melanogaster* has two isoforms of Oskar⁴⁹: Short Oskar, containing the LOTUS, OSK and interdomain regions, and Long Oskar, containing all domains of Short Oskar as well as an additional 5' domain (Figure B.7). It was previously reported that Long Oskar was absent from *N. vitripennis* and *C. pipiens*⁶, and within our alignment of Oskar sequences we could only detect the Long Oskar isoform within Diptera. Therefore, using our dataset, we asked when these two isoforms had evolved. We selected the dipteran sequences from our Oskar alignment and then grouped the sequences by family. We plotted the amino acid occupancy at each alignment position (Figure B.7). We found that Long Oskar predates the Drosophilids, being found as early as the *Pinpunculidae* (Figure B.7). Moreover, following the evolution of the Long Oskar isoform, the Long Oskar domain was conserved in all families except for the *Glossinidae* and *Scathophagidae*. However, given that we found only eight and two Oskar sequences for these families respectively, we cannot eliminate the possibility that apparent absence of the Long Oskar domain in these groups reflects our small sample size, rather than true evolutionary loss.

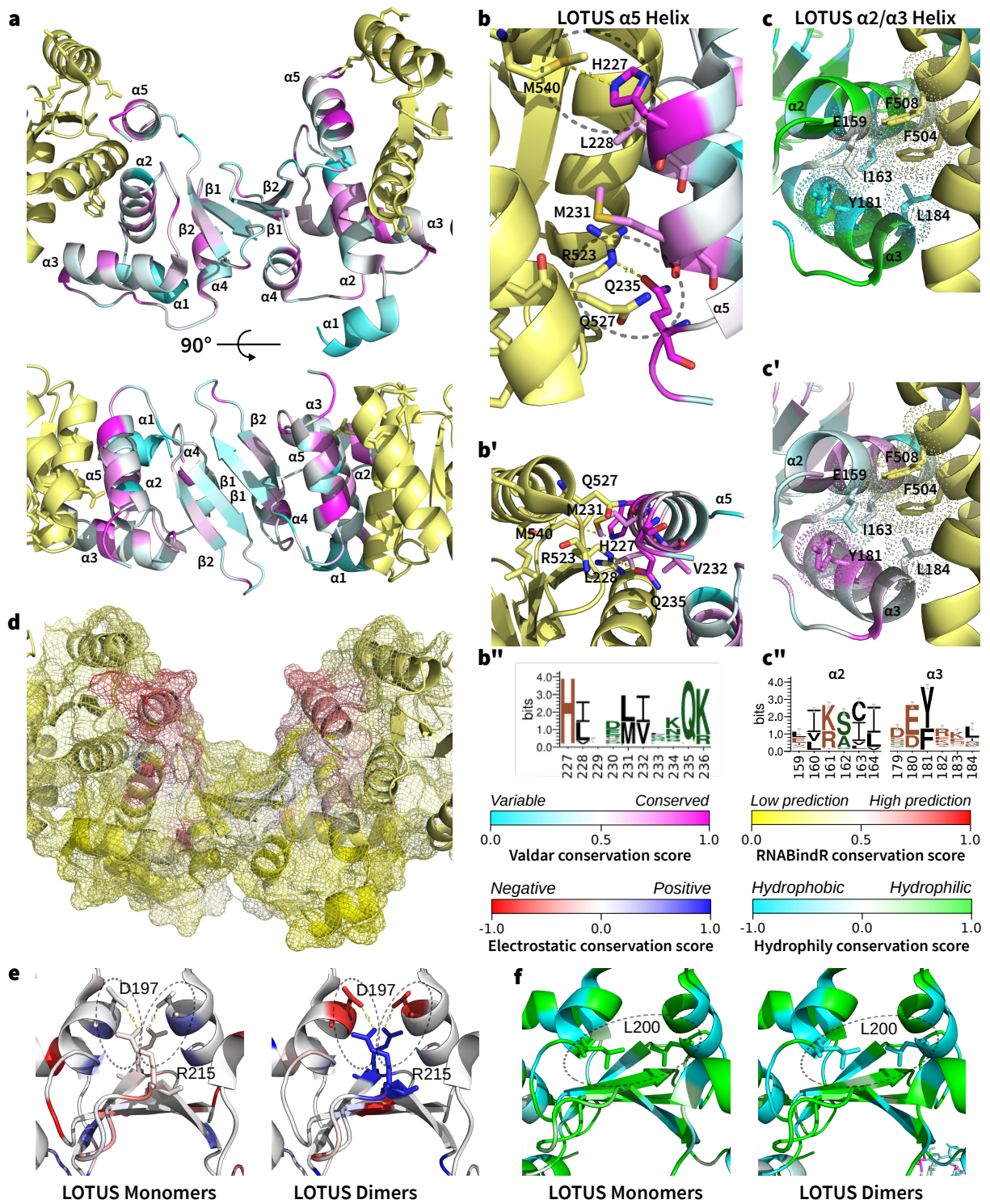
2.3.5 Evidence for evolution of stronger dimerization potential of the Oskar LOTUS domain in Holometabola

The LOTUS domain dimerizes *in vitro* through electrostatic and hydrophobic contacts of Arg215 of the β 2 sheet and Thr195, Asp197 and Leu200 of the α 2 helix^{22,23}. To date, however, the biological significance of Oskar dimerization remains unknown. Moreover, the dimerization of the LOTUS domain does not appear to be conserved across all Oskar sequences²². Specifically, ten LOTUS domains from non-Drosophilid species were reported as tested for dimerization, and only LOTUS domains from *Drosophilidae*, *Tephritidae* and *Pteromalidae* formed homodimers²². The other sequences tested, from *Culicidae*, *Formicidae* and *Gryllidae*, remained monomeric under the tested conditions²². We selected the LOTUS sequences in our alignment from those six families and placed them into one of two groups, dimeric and monomeric LOTUS, under the assumption that any sequence from that family would conserve the dimerization (or absence thereof) properties previously reported²². We asked whether we could detect any evolutionary changes in protein sequence between the two groups, by looking at different properties of known important dimerization interfaces and residues in our sequence alignment²².

In the *D. melanogaster* structure, two key amino acids, D197 and R215, are predicted to form hydrogen bonds that stabilize the dimer²². We found that in the dimer group, the electrostatic properties of these two amino acids are highly conserved (-0.75 for D197 and 0.81 for R215), while in the monomer group the electrostatic interaction is not conserved (0.03 for D197 and -0.11 for R215) (Figure 2.6e). Given the differential conservation between the two groups, our results support the previous finding that disrupting this interaction prevents the dimerization²². L200 was previously hypothesized to stabilize the interface via hydrophobic forces²². We observed that the hydrophobicity of this residue is highly conserved in the dimer group (L200: 0.89), but that in the monomer group this residue is hydrophilic (L200: 2.33) (Figure 2.6f). In sum, our analyses show that key amino acids in the LOTUS domain evolved differently in distinct insect lineages, in a way that may explain why some insect LOTUS domains dimerize and some do not.

Figure 2.6 (following page): Conservation analysis of the LOTUS domain. In **a**), cartoon rendering of the LOTUS domain in complex with Vasa (PDBID: 5NT7) from two different angles. Each amino acid is colored based on its Valdar conservation score. The α helices and β sheets are displayed on top of the structure. In yellow is the Vasa protein. In **b**) and **c**) Logo of the $\alpha 5$ and $\alpha 2$ helix respectively, generated with Weblogo. In black are hydrophobic residues, in blue charged residues, and in green polar residues. In **b, b')**, close up of the conserved $\alpha 5$ helix, with key amino acids displayed as sticks and colored by Valdar conservation score. Two potential contacts are highlighted with dashed lines, namely: H227 and Q235. In **c, c')**, close up of the conserved $\alpha 2$ helix, with key amino acids displayed as sticks and colored by hydrophily conservation score. In **d**), surface mesh rendering colored with the RNABindR RNA binding conservation score. In **e, f**), close up of the LOTUS β sheet dimerization interface. On the left is the conservation in the monomeric LOTUS and on the right in the dimeric LOTUS. In **e**), Amino acids are colored by electrostatic conservation score. Highlighted with dashed lines is the key electrostatic interaction stabilizing the dimerization. In **f**), Amino acids are colored by hydrophobicity conservation scores. Highlighted with dashed lines is the key hydrophobic pocket stabilizing the dimerization.

Figure 2.6: (continued)



2.3.6 Conservation of the Oskar-Vasa interaction interface

Next, we asked whether we could detect differential conservation of the LOTUS-Vasa interface. It was previously reported that the LOTUS domain acts as an interaction domain with Vasa, a key protein in the establishment of the germline^{3,21}. The interaction between LOTUS and Vasa acts through an interaction surface situated in the pocket formed by the helices $\alpha 2$ and $\alpha 5$ of the LOTUS domain (Figure 2.6a b and c). Due to the essential role that *vasa* plays in germline determination (reviewed in Extavour and Akam², Noce et al.⁵⁰, Raz⁵¹, Ewen-Campen et al.⁵²), and the potential co-option of *oskar* to the germline determination mechanism in Holometabola⁷, we hypothesized that changes in the conservation of the residues of this interface might be detectable. First, we observed that the residues of the LOTUS domain $\alpha 2$ and $\alpha 5$ helices contacting Vasa were highly conserved overall ($\alpha 2$ average Valdar score: 0.49 and $\alpha 5$ Valdar score 0.56) (Figure 2.6b). Specifically, we observed that the previously reported Vasa interacting amino acids A162 and L228 of the LOTUS domain were highly conserved (Valdar score: 0.64 for both residues)²¹. We also noted that Q235 and H227 of the LOTUS domain $\alpha 5$ helix are likely to be important interaction partners due to their high conservation (Valdar score: 0.90 and 0.90 respectively) (Figure 2.6b). Moreover, facing the LOTUS domain H227 is Vasa M540, which may act as a proton donor to form a hydrogen bond between the His ring and the sulfur atom of the Met⁵³ (Fig 6b and b'). The LOTUS domain $\alpha 2$ helix is overall slightly less conserved than the LOTUS domain $\alpha 5$ helix (Valdar score: 0.49 vs 0.56) (Figure 2.6a, b", c"), but hydrophobic properties are conserved on one side of the $\alpha 2$ helix (Figure 2.6c, c') forming a motif of conserved amino acid properties (Figure 2.6c").

Previous reports have hypothesized that the *D. melanogaster* LOTUS domain could act as a dsRNA binding domain^{20,54}. However, in *D. melanogaster*, it was later reported that the LOTUS domain did not bind to nucleotides²². Therefore, using our dataset we predicted the RNA binding properties of LOTUS domains to test the conservation of this prediction. We used the RNABindR algorithm³⁹ to predict potential RNA binding sites of the LOTUS domain, and computed a conservation score for each position³⁹. The $\alpha 5$ helix is also the location in the LOTUS domain that has the most conserved prediction for RNA binding (Figure 2.6d).

Finally, we asked whether the secondary structure of the LOTUS domain might be conserved. Secondary structures are often indicative of the tertiary structure of a domain. Therefore, the secondary structure might be conserved even if the sequence varies. We submitted the LOTUS sequences from all identified Oskar orthologs to the Jpred4 servers⁴⁰ for secondary structure prediction and mapped the results onto the Oskar alignment we obtained. The secondary structure of LOTUS is highly conserved throughout Oskar orthologs, with the exception of the α 1 helix (Figure B.8) which also displays a low conservation score of 0.19 (Figure 2.6a).

2.3.7 The core of the OSK domain is conserved

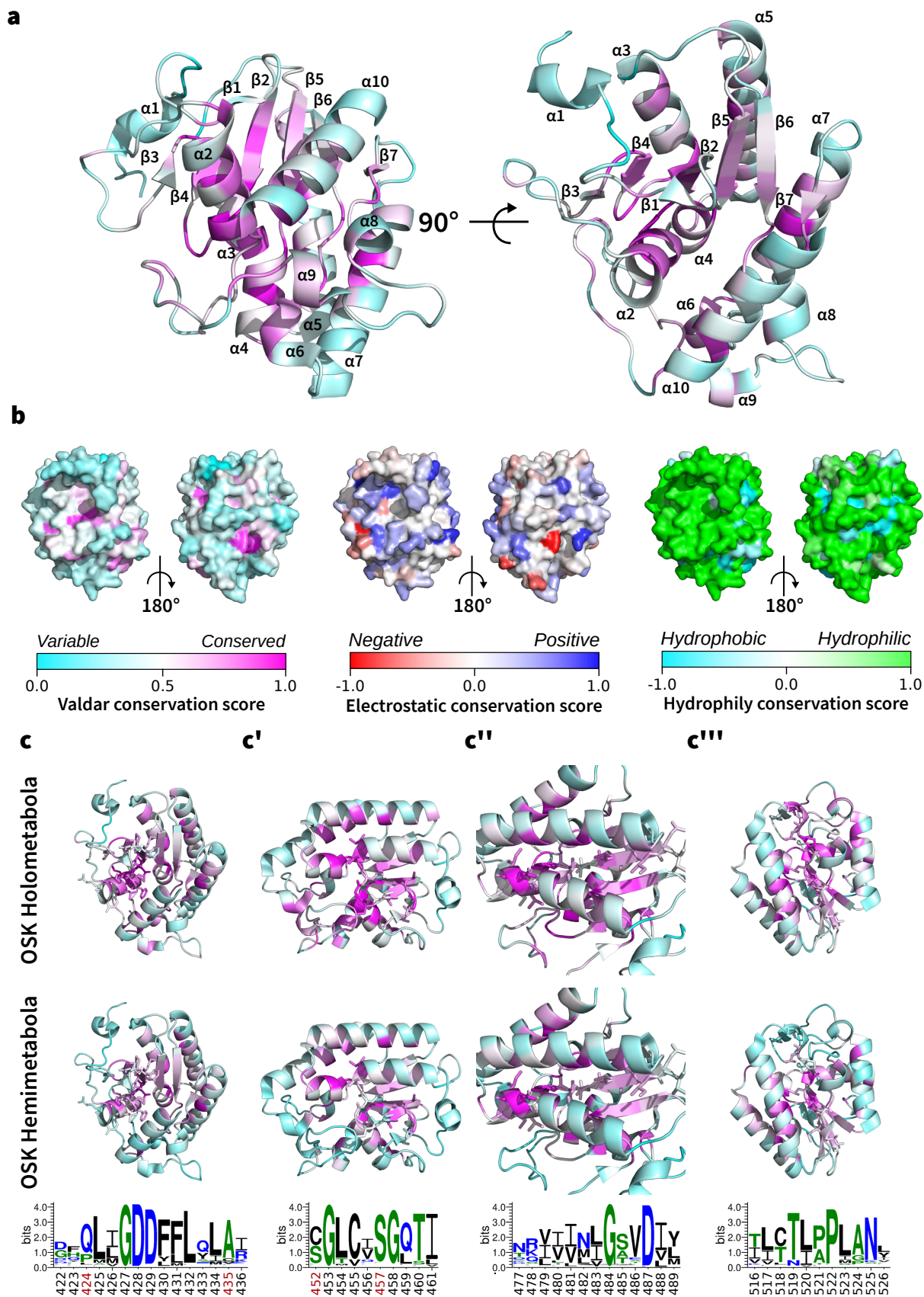
We asked whether the OSK domain showed any differential conservation across the different parts of the domain. We found that the OSK domain of Oskar showed an overall conservation across all insects, similar to the LOTUS domain (Valdar score: 0.51) (Figure 2.7a). However, the conservation pattern is higher in the core amino acids (Valdar score average of core amino acid: 0.54) when compared to the residues at the surface (Valdar score average for surface amino acid: 0.23) (Figure 2.7a). Despite the overall low conservation of the residues at the surface of the OSK domain, we found that the electrostatic properties are conserved overall (electrostatic conservation score > 0; conserved) in the previously reported putative RNA binding pocket²³. However, as previously mentioned, this conservation is stronger in holometabolous sequences (Figure 2.5d). Those results are in accordance with the potential role of OSK as an RNA Binding domain^{22,23}. We also submitted the OSK sequences gathered to the same secondary structure analysis performed on LOTUS. In a similar fashion, the secondary structure of OSK is highly conserved throughout all insect sequences analyzed (Figure B.8).

We then asked if the conservation patterns observed at the core of OSK were clustered in sequence motifs. When we looked at the location of the highly conserved amino acids, we found that the conservation was driven by four well-defined sequence motifs (Figure 2.7c, c', c'', c'''). Given that *oskar* plays different roles in Holometabola and Hemimetabola, we asked whether the conserved OSK motifs showed any difference in conservation between these two groups. Of the four highly conserved OSK core motifs (Figure 2.7c, c', c'', c'''), two of them (Figure 2.7c Valdar average score: 0.80 and c'' Valdar average score: 0.71) were conserved across all insects, but the

other two showed differential conservation between the holometabolous and hemimetabolous sequences (Figure 2.7c' Valdar score average holo: 0.78 hemi: 0.58 and c''' Valdar score average holo: 0.70 hemi: 0.55). Finally, we noted that only one of the affected residues in known alleles affecting posterior patterning in *D. melanogaster*, S457, is conserved across all insects (Valdar score: 0.86). This suggests that the role of the other previously reported important amino acids in the function of *D. melanogaster* OSK²³ might not be conserved in other insects (red positions in Figure 2.7c, c', c'', c''').

Figure 2.7 (following page): Conservation analysis of the OSK domain. In **a**), cartoon rendering of the OSK domain (PDBID: 5A4A) from two different angles. Each amino acid is colored based on its Valdar conservation score. Shown in sticks are known amino acid location of *D. melanogaster* alleles leading to the loss of *oskar* polarization. In **b**), surface rendering of the OSK domain colored by Valdar conservation score, electrostatic conservation score and hydrophobicity conservation score. In **c**, **c'**, **c''**, **c'''**), close up of conserved motifs of the OSK domain. Each location is colored with Valdar conservation scores of the Hemimetabola and Holometabola sequences. At the bottom are sequence logos of each motif generated with Weblogo. In black are hydrophobic residues, in blue charged residues, and in green polar residues. Positions in red are amino acid locations of *D. melanogaster* alleles leading to the loss of *oskar* polarization. Shown in sticks in the structure rendering are the amino acids displayed in the logo.

Figure 2.7: (continued)



2.4 Discussion

2.4.1 An expanded collection of *oskar* orthologs

oskar provides a powerful case study of functional evolution of a gene with an unusual genesis⁸. Here, we gathered the most extensive set of orthologous *oskar* sequences to date. However, most insect genomic and transcriptomic data have been generated from only a few orders, and the vast majority from the Holometabola. Diptera, Lepidoptera, Coleoptera, Hymenoptera and Hemiptera represent 82% of the available datasets. We emphasize that expanded taxon sampling, particularly for the Hemimetabola, will be critical for further studies of the evolution of protein function across insects. Moreover, only a fraction (27% for tissue type, 26% for organism stage, and 14% for sex) of the TSA datasets contained usable metadata regarding the stage and tissue type sampled. Future standardization of the nature and format of transcriptomic metadata would also be a worthwhile endeavor that could increase the efficiency and efficacy of future work.

2.4.2 Convergent losses of *oskar* in insect evolution

Our observations strongly suggest the loss of *oskar* in the Lepidoptera. This is supported by the fact that a common ancestor of Lepidopteran and other holometabolous insects had already established the new, *oskar* dependent, inheritance germline specification (See Figure 2.3 for phylogenetic relationships). In the wasp *Nasonia vitripennis*, a Hymenopteran, *oskar* is known to be expressed at the posterior of the oocyte and in germ cells⁶ as well as being a necessary component of the function of the Oosome (the wasp germ plasm homolog)⁵⁵. Therefore, the apparent subsequent loss in nearly all Lepidoptera examined, of a gene responsible for the establishment of the germ plasm in other Holometabola, was unexpected (but note the previously reported absence of *oskar* from the *Bombyx mori* genome⁶). This suggests either that Lepidopteran species that lack *oskar* do not use germ plasm to specify their germ line, or that they do so using a germ plasm nucleator other than Oskar. As was previously discussed in Quan and Lynch²⁵, Lepidopteran species such as *Bombyx mori*⁵⁶ or *Parage aegeria*⁵⁷ seem to harbor a

germ plasm specification strategy.

2.4.3 Functional implications of differential conservation of the LOTUS and OSK domains

We have identified novel conserved amino acid positions that we hypothesize are important for the LOTUS Vasa binding and RNA binding of OSK (Figures 2.6 and 2.7). Our observation of the conservation of the $\alpha 2$ helix is consistent with its previously reported importance in the LOTUS domain's capacity to interact with Vasa²¹. In the $\alpha 2$ helix, we also observed high conservation of H227 and Q235, whose position suggests they may contribute to the interaction between Vasa and LOTUS, and that we suggest should be the target of future mutational studies.

We also uncovered an interesting new conservation pattern within the OSK domain. The conserved amino acids were higher in the core of the domain than on the surface. This differential conservation might reflect the acquisition of the germ plasm nucleator role of *oskar* (Figure 2.5). We noted that the basic properties of surface residues previously reported for *D. melanogaster*²³, are conserved across insects. If the RNA binding properties of OSK observed in *D. melanogaster*^{22,23} are conserved throughout holometabolous insects, then it is possible that the low amino acid conservation of the surface, which nevertheless displays highly conserved basic properties, could be due to the co-evolution of specific RNA binding partners for the OSK domains of different lineages.

2.4.4 OSK evolved differentially between holometabolous and hemimetabolous insects

Finally, we observed a differential conservation of the OSK domain between hemimetabolous and holometabolous insects. We found that the OSK sequence was less conserved across the Holometabola than across the Hemimetabola. This observation raises two important and interesting questions about the role of the OSK domain in the functional evolution of Oskar: first, was the apparently relaxed purifying selection experienced by OSK in the Holometabola necessary for the co-option of *oskar* in germ plasm nucleation? Second, is there a function of

Oskar in the hemimetabolous insects that requires strong conservation of OSK? More studies on the roles and biochemical properties of OSK in hemimetabolous insects should be undertaken to further our understanding of the difference in conservation.

In conclusion, analysis of the large dataset of novel Oskar sequences presented here provides multiple new hypotheses concerning the molecular mechanisms and functional evolution of the Oskar gene, that we hope will be tested in the future.

References

- [1] D. L. Kirk. A twelve-step program for evolving multicellularity and a division of labor. *Bioessays*, 27(3):299–310, March 2005.
- [2] C. G. Extavour and M. Akam. Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development*, 130(24):5869–5884, December 2003.
- [3] A. Ephrussi and R. Lehmann. Induction of germ cell formation by *oskar*. *Nature*, 358(6385):387–392, July 1992.
- [4] R. Lehmann and C. Nüsslein-Volhard. Abdominal segmentation, pole cell formation, and embryonic polarity require the localized activity of *oskar*, a maternal gene in *Drosophila*. *Cell*, 47(1):141–152, October 1986.
- [5] J. Juhn and A. A. James. *oskar* gene expression in the vector mosquitoes, *Anopheles gambiae* and *Aedes aegypti*. *Insect Mol. Biol.*, 15(3):363–372, June 2006.
- [6] J. A. Lynch, O. Ozüak, A. Khila, E. Abouheif, C. Desplan, and S. Roth. The phylogenetic origin of *oskar* coincided with the origin of maternally provisioned germ plasm and pole cells at the base of the Holometabola. *PLoS Genet.*, 7(4):e1002029, April 2011.
- [7] B. Ewen-Campen, J. R. Srouji, E. E. Schwager, and C. G. Extavour. Oskar predates the evolution of germ plasm in insects. *Curr. Biol.*, 22(23):2278–2283, December 2012.
- [8] L. Blondel, T. E. Jones, and C. G. Extavour. Bacterial contribution to genesis of the novel germ line determinant *oskar*. *Elife*, 9:e45539, February 2020.

- [9] B. Misof, S. Liu, K. Meusemann, R. S. Peters, A. Donath, C. Mayer, P. B. Frandsen, J. Ware, T. Flouri, R. G. Beutel, O. Niehuis, M. Petersen, F. Izquierdo-Carrasco, T. Wappler, J. Rust, A. J. Aberer, U. Aspöck, H. Aspöck, D. Bartel, A. Blanke, S. Berger, A. Böhm, T. R. Buckley, B. Calcott, J. Chen, F. Friedrich, M. Fukui, M. Fujita, C. Greve, P. Grobe, S. Gu, Y. Huang, L. S. Jermiin, A. Y. Kawahara, L. Krogmann, M. Kubiak, R. Lanfear, H. Letsch, Y. Li, Z. Li, J. Li, H. Lu, R. Machida, Y. Mashimo, P. Kapli, D. D. McKenna, G. Meng, Y. Nakagaki, J. L. Navarrete-Heredia, M. Ott, Y. Ou, G. Pass, L. Podsiadlowski, H. Pohl, B. M. von Reumont, K. Schütte, K. Sekiya, S. Shimizu, A. Slipinski, A. Stamatakis, W. Song, X. Su, N. U. Szucsich, M. Tan, X. Tan, M. Tang, J. Tang, G. Timelthaler, S. Tomizuka, M. Trautwein, X. Tong, T. Uchifune, M. G. Walz, B. M. Wiegmann, J. Wilbrandt, B. Wipfler, T. K. F. Wong, Q. Wu, G. Wu, Y. Xie, S. Yang, Q. Yang, D. K. Yeates, K. Yoshizawa, Q. Zhang, R. Zhang, W. Zhang, Y. Zhang, J. Zhao, C. Zhou, L. Zhou, T. Ziesmann, S. Zou, Y. Li, X. Xu, Y. Zhang, H. Yang, J. Wang, J. Wang, K. M. Kjer, and X. Zhou. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763–767, November 2014.
- [10] P. K. Dearden. Germ cell development in the Honeybee (*Apis mellifera*); vasa and nanos expression. *BMC Dev. Biol.*, 6:6, February 2006.
- [11] L. Nagy, L. Riddiford, and K. Kiguchi. Morphogenesis in the early embryo of the lepidopteran *Bombyx mori*. *Dev. Biol.*, 165(1):137–151, September 1994.
- [12] R. Schröder. vasa mRNA accumulates at the posterior pole during blastoderm formation in the flour beetle *Tribolium castaneum*. *Dev. Genes Evol.*, 216(5):277–283, May 2006.
- [13] J. A. Nelson. *The embryology of the honey bee*. Princeton University Press, 1915.
- [14] J. Kim-Ha, J. L. Smith, and P. M. Macdonald. *oskar* mRNA is localized to the posterior pole of the *Drosophila* oocyte. *Cell*, 66(1):23–35, July 1991.
- [15] A. M. Rafiqi, A. Rajakumar, and E. Abouheif. Origin and elaboration of a major evolutionary transition in individuality. *Nature*, 585(7824):239–244, 2020.
- [16] P. J. Webster, J. Suen, and P. M. Macdonald. *Drosophila virilis oskar* transgenes direct body patterning but not pole cell formation or maintenance of mRNA localization in *D.*

- melanogaster. *Development*, 120(7):2027–2037, July 1994.
- [17] J. R. Jones and P. M. Macdonald. Oskar controls morphology of polar granules and nuclear bodies in *Drosophila*. *Development*, 134(2):233–236, January 2007.
- [18] T. Sikosek, H. S. Chan, and E. Bornberg-Bauer. Escape from Adaptive Conflict follows from weak functional trade-offs and mutational robustness. *Proc. Natl. Acad. Sci. U. S. A.*, 109(37):14888–14893, September 2012.
- [19] T. Sikosek and H. S. Chan. Biophysics of protein evolution and evolutionary protein biophysics. *J. R. Soc. Interface*, 11(100):20140419, November 2014.
- [20] V. Anantharaman, D. Zhang, and L. Aravind. OST-HTH: a novel predicted RNA-binding domain. *Biology Direct*, 5(1):13, 2010.
- [21] M. Jeske, C. W. Müller, and A. Ephrussi. The LOTUS domain is a conserved DEAD-box RNA helicase regulator essential for the recruitment of Vasa to the germ plasm and nuage. *Genes Dev.*, 31(9):939–952, May 2017.
- [22] M. Jeske, M. Bordi, S. Glatt, S. Müller, V. Rybin, C. W. Müller, and A. Ephrussi. The Crystal Structure of the *Drosophila* Germline Inducer Oskar Identifies Two Domains with Distinct Vasa Helicase- and RNA-Binding Activities. *Cell Reports*, 12(4):587–598, 2015.
- [23] N. Yang, Z. Yu, M. Hu, M. Wang, R. Lehmann, and R.-M. Xu. Structure of *Drosophila* Oskar reveals a novel RNA binding protein. *Proc. Natl. Acad. Sci. U. S. A.*, 112(37):11541–11546, September 2015.
- [24] A. Ahuja and C. G. Extavour. Patterns of molecular evolution of the germ line specification gene *oskar* suggest that a novel domain may contribute to functional divergence in *Drosophila*. *Development Genes and Evolution*, 224(2):65–77, 2014.
- [25] H. Quan and J. A. Lynch. The evolution of insect germline specification strategies. *Curr Opin Insect Sci*, 13:99–105, February 2016.
- [26] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, 34(Web Server issue):

- W435–9, July 2006.
- [27] F. Madeira, Y. M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A. R. N. Tivey, S. C. Potter, R. D. Finn, and Others. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, 47(W1):W636–W641, 2019.
- [28] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, March 2004.
- [29] R. W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, April 1950.
- [30] S. Capella-Gutierrez, J. M. Silla-Martinez, and T. Gabaldon. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, 2009.
- [31] A. Rausell, D. Juan, F. Pazos, and A. Valencia. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl. Acad. Sci. U. S. A.*, 107(5):1995–2000, February 2010.
- [32] A. Löytynoja. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.*, 1079:155–170, 2014.
- [33] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.
- [34] J. Huerta-Cepas, F. Serra, and P. Bork. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.*, 33(6):1635–1638, June 2016.
- [35] W. S. J. Valdar. Scoring residue conservation. *Proteins*, 48(2):227–241, August 2002.
- [36] J. A. Capra and M. Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, August 2007.
- [37] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, 37(1):145–151, January 1991.

- [38] C. P. Moon and K. G. Fleming. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proc. Natl. Acad. Sci. U. S. A.*, 108(25): 10174–10177, June 2011.
- [39] M. Terribilini, J. D. Sander, J.-H. Lee, P. Zaback, R. L. Jernigan, V. Honavar, and D. Dobbs. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.*, 35(Web Server issue):W578–84, July 2007.
- [40] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.*, 43(W1):W389–94, July 2015.
- [41] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14(6):1188–1190, June 2004.
- [42] W. L. DeLano and Others. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography*, 40(1):82–92, 2002.
- [43] C. Mitter, D. R. Davis, and M. P. Cummings. Phylogeny and Evolution of Lepidoptera. *Annu. Rev. Entomol.*, 62:265–283, January 2017.
- [44] A. Y. Kawahara, D. Plotkin, M. Espeland, K. Meusemann, E. F. A. Toussaint, A. Donath, F. Gimnich, P. B. Frandsen, A. Zwick, M. Dos Reis, J. R. Barber, R. S. Peters, S. Liu, X. Zhou, C. Mayer, L. Podsiadlowski, C. Storer, J. E. Yack, B. Misof, and J. W. Breinholt. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl. Acad. Sci. U. S. A.*, 116(45):22657–22663, November 2019.
- [45] R. S. Peters, L. Krogmann, C. Mayer, A. Donath, S. Gunkel, K. Meusemann, A. Kozlov, L. Podsiadlowski, M. Petersen, R. Lanfear, P. A. Diez, J. Heraty, K. M. Kjer, S. Klopstein, R. Meier, C. Polidori, T. Schmitt, S. Liu, X. Zhou, T. Wappler, J. Rust, B. Misof, and O. Niehuis. Evolutionary History of the Hymenoptera. *Curr. Biol.*, 27(7):1013–1018, April 2017.
- [46] J. Dubnau, A.-S. Chiang, L. Grady, J. Barditch, S. Gossweiler, J. McNeil, P. Smith, F. Buldoc, R. Scott, U. Certa, C. Broger, and T. Tully. The staufer/pumilio pathway is involved in *Drosophila* long-term memory. *Curr. Biol.*, 13(4):286–296, February 2003.

- [47] B. Ewen-Campen, S. Donoughe, D. N. Clarke, and C. G. Extavour. Germ cell specification requires zygotic mechanisms rather than germ plasm in a basally branching insect. *Curr. Biol.*, 23(10):835–842, May 2013.
- [48] L. Lebart, A. Morineau, and K. M. Warwick. *Multivariate descriptive statistical analysis: Correspondence analysis and related techniques for large matrices*. John Wiley & Sons; Chichester, UK, 1984.
- [49] F. H. Markussen, A. M. Michon, W. Breitwieser, and A. Ephrussi. Translational control of *oskar* generates short OSK, the isoform that induces pole plasm assembly. *Development*, 121(11):3723–3732, November 1995.
- [50] T. Noce, S. Okamoto-Ito, and N. Tsunekawa. Vasa homolog genes in mammalian germ cell development. *Cell Struct. Funct.*, 26(3):131–136, June 2001.
- [51] E. Raz. The function and regulation of vasa-like genes in germ-cell development. *Genome Biol.*, 1(3):REVIEWS1017, September 2000.
- [52] B. Ewen-Campen, E. E. Schwager, and C. G. M. Extavour. The molecular machinery of germ line specification. *Mol. Reprod. Dev.*, 77(1):3–18, 2010.
- [53] D. Pal and P. Chakrabarti. Non-hydrogen bond interactions involving the methionine sulfur atom. *J. Biomol. Struct. Dyn.*, 19(1):115–128, August 2001.
- [54] I. Callebaut and J.-P. Mornon. LOTUS, a new domain associated with small RNA pathways in the germline. *Bioinformatics*, 26(9):1140–1144, 2010.
- [55] H. Quan, D. Arsala, and J. A. Lynch. Transcriptomic and functional analysis of the oosome, a unique form of germ plasm in the wasp *Nasonia vitripennis*. *BMC Biol.*, 17(1):78, October 2019.
- [56] H. Nakao, T. Matsumoto, Y. Oba, T. Niimi, and T. Yaginuma. Germ cell specification and early embryonic patterning in *Bombyx mori* as revealed by nanos orthologues. *Evol. Dev.*, 10(5):546–554, September 2008.

- [57] J.-M. Carter, M. Gibbs, and C. J. Breuker. Divergent RNA Localisation Patterns of Maternal Genes Regulating Embryonic Patterning in the Butterfly *Pararge aegeria*. *PLoS One*, 10(12): e0144471, December 2015.

No observations on single genes can ever illuminate the overall mechanisms of development of the body plan or of body parts except at the minute and always partial, if not wholly illusory, level of the worm's eye view. The same must be true as well of major evolutionary change in the body plan or in body parts.

Eric H. Davidson, 2011

3

Topology-driven protein-protein interaction
network analysis detects genetic sub-networks
regulating reproductive capacity

ABSTRACT

Understanding the genetic regulation of organ structure is a fundamental problem in developmental biology. Here, we use egg-producing structures of insect ovaries, called ovarioles, to deduce systems-level gene regulatory relationships from quantitative functional genetic analysis. We previously showed that Hippo signalling, a conserved regulator of animal organ size, regulates ovariole number in *Drosophila melanogaster*. To comprehensively determine how Hippo signalling interacts with other pathways in this regulation, we screened all known signalling pathway genes, and identified Hpo-dependent and Hpo-independent signalling requirements. Network analysis of known protein-protein interactions among screen results identified independent gene regulatory sub-networks regulating one or both of ovariole number and egg laying. These sub-networks predict involvement of previously uncharacterised genes with higher accuracy than the original candidate screen. This shows that network analysis combining functional genetic and large-scale interaction data can predict function of novel genes regulating development.

CONTRIBUTIONS

This work was published in the journal *Elife* in 2020 and has been reformatted as the chapter presented here:

Kumar, T.*, Blondel, L.*, & Extavour, C. G. (2020). **Topology-driven protein-protein interaction network analysis detects genetic sub-networks regulating reproductive capacity.** *ELife*, 9.

DOI: [10.7554/elife.54082](https://doi.org/10.7554/elife.54082)

The work presented here is the result of the collaboration between Tarun Kumar, Cassandra G. Extavour and myself. Tarun Kumar and Cassandra G. Extavour designed the screen, and Cassandra G. Extavour directed the study design, funding, writing and reviewing. Tarun Kumar performed all the biological experiments. He collected egg counts and ovariole numbers for each of the *Drosophila melanogaster* lines used in this study. Finally, he wrote a major part of the manuscript. I, Leo Blondel, proposed and designed the data analysis. I created all the scripts and performed all the data analysis presented here. I assisted Tarun Kumar on the writing of the manuscript. I created all the figures presented in the manuscript.

ACKNOWLEDGEMENT

I cannot stress enough that this study is the fruit of the amazing collaboration between Tarun Kumar, Cassandra G. Extavour and I. I had the honor of working with Tarun Kumar, a really brilliant scientist with whom long work hours always turned into exciting productive time. It is important for me to thank him prior to this chapter for his skill, his dedication and his patience. I also want to thank Cassandra G. Extavour for the trust she placed in both of us to complete this project and the pertinent advices she always delivered. Finally, I would like to thank Marc Santolini for the many inputs and advises given towards bettering the network science part of this chapter. His help has be invaluable.

3.1 Introduction

The final shape and size of an organ is critical to organismal function and viability. Defects in human organ morphology cause a multitude of pathologies, including cancers, organ hypertrophies and atrophies (e.g. Yang and Xu¹). It is thus critical to understand the regulatory mechanisms underlying the stereotypic shape and size of organs. To this end, assessing the genetic regulation of size is significantly facilitated by using quantifiable changes in organ size and shape.

The *Drosophila melanogaster* female reproductive system is a useful paradigm to study quantitative anatomical traits. In these organs, the effects of multiple genes and the environment combine to produce a quantitative phenotype: a species-specific average number of egg-producing ovarian tubes called ovarioles. Fruit fly ovaries can contain as few as one and as many as 50 ovarioles per ovary, depending on the species^{2,3,4,5}, with each ovariole capable of producing eggs. Ovariole number, therefore, may affect the reproductive fitness of *Drosophila* species by determining the potential of an adult female to produce eggs^{6,7}. While ovariole number within a species can vary across temperatures⁸, altitudinal and latitudinal clines^{9,10}, under constant environmental conditions ovariole number is highly stereotypic^{7,11,12,13}. The reproducibility of ovariole number thus indicates a strong genetic component⁵. Genome wide association studies and quantitative trait locus mapping have demonstrated that the ovariole number is a highly polygenic trait^{14,15,16,17,18,19}. In contrast, functional genetic studies have identified only a small number of genes whose activity regulates ovariole number (discussed below). Thus, the complexity of the genetic regulation of this important trait remains largely unknown.

The determination of ovariole number in *D. melanogaster* occurs during late larval and pupal development²⁰. Each ovariole in the adult fly arises from a single primordial structure called a terminal filament (TF), which forms in the late third instar larval ovary²¹ by convergent extension²² of the terminal filament cells (TFCs)^{21,23}. TFCs are first specified from an anterior population of somatic cells in the larval ovary by the expression of transcription factors including Bric-à-brac 1/2 (*bric-à-brac 1/2*; *bab1/2*) and Engrailed (*engrailed*; *en*)^{21,24}. Initially a

loosely arranged group in the anterior of the larval ovary, TFCs undergo morphogenetic movements to give rise to the ordered columns of cells that are TFs. Cell intercalation during convergent extension is dependent on the actin regulators Cofilin (*twinstar*) and the large Maf factor Traffic Jam (*traffic jam; tj*), and on E-cadherin dependent adhesion^{21,25}. Regulation of ovariole number is thus largely dependent on the specification of the TFCs and their rearrangement into TFs²⁶.

We previously showed that the regulation of both TFC and TF number is dependent on the Hippo signalling pathway²⁶, a pan-metazoan regulator of organ and tissue size^{27,28}. At the core of the Hippo kinase cascade are two protein kinases, Hippo (*hippo; hpo*) and Warts (*warts*), which prevent the nuclear localisation of the transcriptional co-activator Yorkie (*yorkie; yki*). Yki and the transcription factor Scalloped (*scalloped*) together initiate the transcription of multiple target genes, including those that promote cell proliferation and survival. In the *D. melanogaster* larval ovary, loss of Hpo in the somatic cells causes an increase in nuclear Yki, leading to an increase in TFCs, TFs, ovariole number and egg laying in adults²⁶.

Production of fertile eggs from a stereotypic number of ovarioles requires a spatially and temporally coordinated interplay of signalling between the somatic and germ line cells of the ovary. Thus, signalling amongst somatic and germ line cells in the larval ovary is crucial to all stages of ovarian development^{26,29,30,31,32,33}. For instance, disruptions in insulin or Tor signalling affect both somatic and germ line cell proliferation^{26,32,33,34,35,36}. Similarly, ecdysone pulses from the prothoracic gland regulate the timely differentiation of the primordial germ cells (PGCs) and the somatic TFCs^{37,38,39}. Both Hpo and ecdysone signalling also control the proportion of germ line to somatic cells by differentially regulating proliferation of both cell types^{26,37}.

Although it is clear that genes function together in regulatory networks⁴⁰, determining how the few genes functionally verified as required for ovariole development and function, work together to coordinate ovariole number and ovarian function more generally, is a challenge because most genes or pathways have been considered individually. An alternative approach that is less often applied to animal developmental genetics, is a systems biology representation of complex biological systems as networks^{41,42}. Protein-protein interaction networks (PINs) are

such an example⁴³. The availability of high throughput molecular biology datasets from, for example, yeast two-hybrid, protein ChIP and microarrays has allowed for the emergence of large scale interaction networks representing both functional and physical molecular interactions^{40,44,45,46}.

With ample evidence that signalling in the ovary can affect ovarian development, but few genes functionally verified to date, we aimed to identify novel regulators of ovariole development by functionally testing all known members of all characterized *D. melanogaster* signalling pathways. We used tissue-specific RNAi to systematically knock down 463 genes in the larval ovary and looked for modifiers of the *hpo* loss of function egg laying and ovariole number phenotypes. To analyse the results of this phenotypic analysis, we used topology-driven network analysis to identify genetic networks regulating these phenotypes, thus generating hypotheses about the relationships between these networks. With this systems biology approach, we identify not only signalling pathway genes, but also previously untested genes that affect these reproductive traits. Functional testing showed that these novel genes affect ovariole number and/or egg laying, providing us with a novel *in silico* method to identify target genes that affect ovarian development and function. We use these findings to propose putative developmental regulatory networks underlying one or both of ovariole formation and egg laying.

3.2 Results

3.2.1 An RNAi modifier screen for signalling pathway involvement in ovariole number

To systematically ascertain the function of signalling pathway genes and their interactions with Hippo signalling in the development of the *D. melanogaster* ovary, we first curated a list of all known and predicted signalling genes^{47,48,49}. We identified 475 genes belonging to the 14 developmental signalling pathways characterised in *D. melanogaster* (Table 3.1; Table C.1), and obtained UAS:RNAi lines for 463 of these genes from the Vienna *Drosophila* RNAi centre (VDRC) or the TRiP collections at the Bloomington *Drosophila* Stock centre (BDSC) (all *D. melanogaster*

Signalling pathway	Number of genes in screen
EGF	45
FGF	25
FOXO	67
Hippo	60
JAK/STAT	31
JNK	28
MAPK	29
Notch	48
SHH	54
TGF B	52
Toll	36
VEGF	17
Wnt	125
mTOR	36

Table 3.1: Number of candidate genes tested in each signalling pathway. Candidate genes are grouped by their reported roles in one or more signalling pathways based on published literature. Genes in this list are not necessarily unique to a single pathway, but rather may function in more than one signalling pathway. The list of specific genes per pathway that were included in the screen for functional analysis (Figure 3.1) is found in the Table C.1.

genetic lines used are listed in Methods).

We previously showed that reducing the levels of *hpo* in the somatic cells of the larval ovary using *traffic jam Gal4* (*tj:Gal4*) driving *hpo[RNAi]* increased both ovariole number and egg laying of adult female flies²⁶. To identify genes that modify these phenotypes, we used *tj:Gal4* to drive simultaneous *hpo[RNAi]* and *RNAi* against a signalling candidate gene, and quantified the phenotypic change (Figure 3.1a-d). We observed that on driving two copies of *hpo[RNAi]* using *tj:Gal4*, we obtained a further increase in both egg laying and ovariole number (Figure 3.1e). This indicates that ovaries have further potential to increase ovariole number and egg laying

beyond the increase induced by *tj:Gal4* driving one copy of *hpo[RNAi]*, and that *tj:Gal4* can drive the expression of two RNAi constructs, indicating that our screen could identify both enhancers and suppressors of the *tj:Gal4>hpo[RNAi]* phenotype.

We proceeded to identify modifiers of the *tj:Gal4>hpo[RNAi]* phenotype by crossing males of each of the 463 candidate genes RNAis individually with *tj:Gal4>hpo[RNAi]* females, and performing three phenotypic screens on the offspring. In the first screen (Figure 3.1a), we measured egg laying of three F1 female offspring (*tj:Gal4>hpo[RNAi]*, *signalling candidate[RNAi]*) over 5 days. To address batch variation (Figure C.1), we standardized egg laying measurements by calculating the Z scores (Z_{gene} = number of standard deviations from the mean) for each candidate line relative to its batch controls. 190 genes had an egg laying $|Z_{gene}|$ below 1. Previous studies have shown that the egg laying of newly eclosed adult mated females correlates with ovariole number during the first five days⁶. We therefore eliminated these 190 genes from subsequent screening, because the change in egg laying was so modest that we considered these candidates were unlikely to show changes in ovariole number when compared to controls.

In the second screen (Figure 3.1b), we measured egg laying in a wild-type background (*tj>signalling candidate[RNAi]*) for the 273 remaining candidate genes. For the third screen (Figure 3.1c), we quantified the ovariole number of *tj:Gal4>hpo[RNAi]*, *signalling candidate[RNAi]* F1 adult females for the same 273 candidate genes. To choose candidates from the second and third screens for further study, we wished to account for the fact that the two screens had different effective numbers of data points. This was because egg laying data were obtained from individual vials of three females over five days, while ovariole numbers were obtained from 20 ovaries from ten females (see methods). We therefore selected the 67 genes with a $|Z_{gene}|$ above two for ovariole number (Figure 3.1c, d; Table 3.2), and the 49 genes with a more conservative $|Z_{gene}|$ above five for egg laying (Figure 3.1a, b, d; Table 3.2), for a total of 116 positive candidates for subsequent analyses.

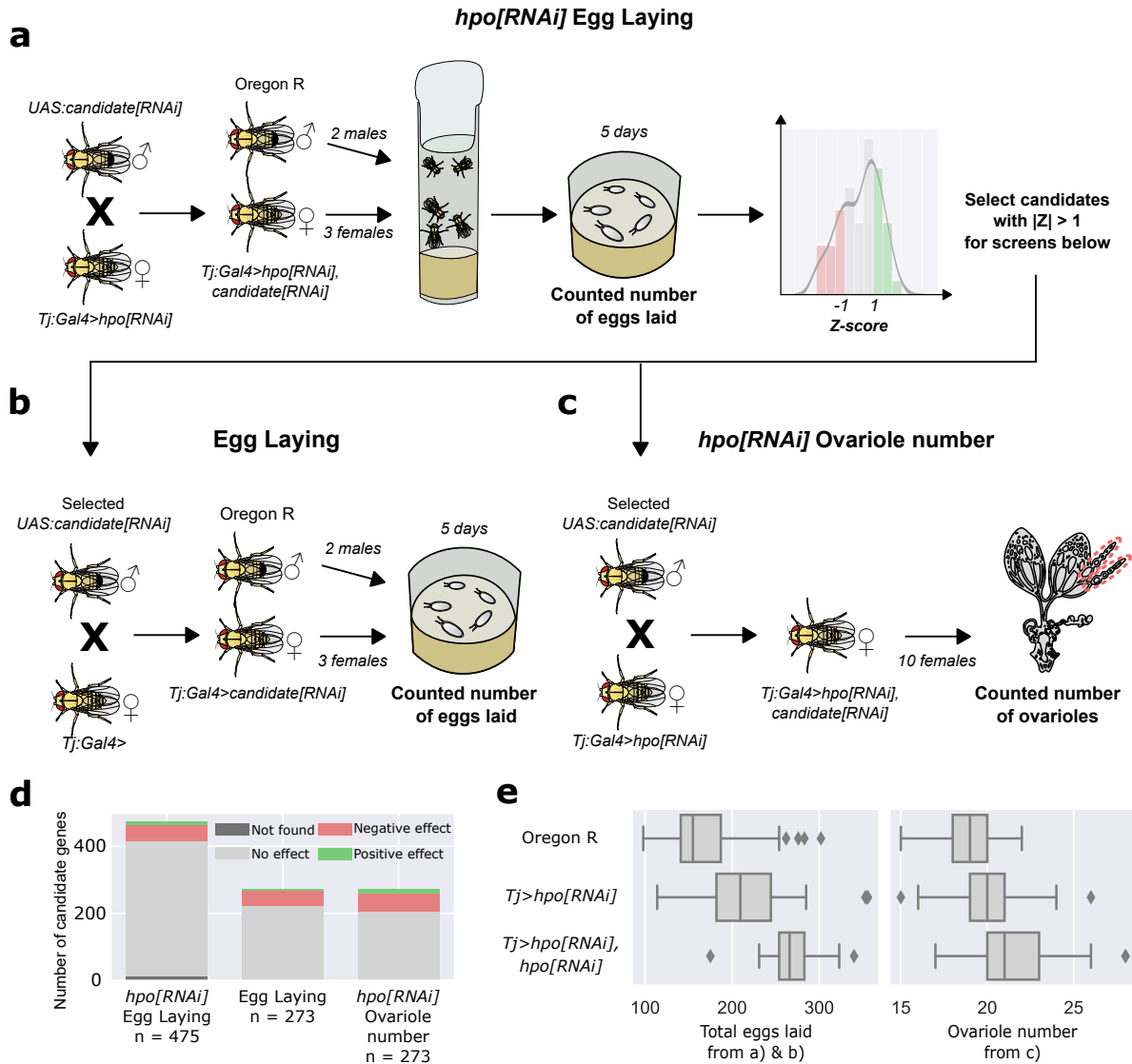


Figure 3.1: Screen methodology. **a,b,c**) Diagrammatic representation of screen workflow. **d**) Distributions of results of genes in the three screens. n = number of genes tested in each screen (see also Table 3.2). **e**) Total eggs laid by three female flies over five days (left panel) and ovariole number (right panel) of Oregon R (top row), *tj:Gal4* driving one copy of *UAS:hpo*[RNAi] (middle row), and *tj:Gal4* driving two copies of *UAS:hpo*[RNAi] (bottom row), showing that, the previously reported *tj:Gal4>hpo*[RNAi] ovariole number and egg laying phenotypes²⁶ can be modified by further UAS:RNAi-mediated gene knockdown. Distribution of egg laying and ovariole number of controls in each screen batch is illustrated in **Figure C.1**.

Egg Laying Screens	<i>hpo</i>[RNAi] Egg Laying (Figure 1a)	Egg Laying (Figure 1b)	Ovariole Number Screen	<i>hpo</i>[RNAi] Ovariole Number (Figure 1c)
RNAi stocks unavailable	12	0	RNAi stocks unavailable	0
Primary filter ($ Z_{gene} < 1$)	190	N/A	Primary filter ($ Z_{gene} < 1$)	N/A
No effect ($-5 < Z_{gene} < 5$)	214	224	No effect ($-2 < Z_{gene} < 2$)	206
Negative effect ($Z_{gene} < -5$)	48	44	Negative effect ($Z_{gene} < -2$)	54
Positive effect ($Z_{gene} > 5$)	11	5	Positive effect ($Z_{gene} > 2$)	13
Total	475	273	Total	273

Table 3.2: Results of the three functional genetic screens. Number of genes tested in each screen and cumulative results. "Negative effect" corresponds to a reduction in eggs laid or number of ovarioles below the Z score (Z_{gene}) threshold for each phenotype. "Positive effect" indicates an increase above the set Z_{gene} thresholds. Z_{gene} thresholds for each category in each screen are indicated in brackets. The primary filter of $|Z_{gene}| < 1$ was applied only to the *hpo*[RNAi] Egg Laying screen shown in Figure 3.1a. The list of specific genes that exceeded our chosen Z_{gene} thresholds for each scored phenotype (Figure 3.1) and were therefore considered to have a positive or negative effect on the phenotype, is found in the Table C.1. The 12 genes for which RNAi stocks were unavailable at the time of testing are listed in Table 3.3.

FbID	CG Number	Name	Symbol
FBgn0283468	CG3412	<i>supernumerary limbs</i>	<i>slmb</i>
FBgn0267821	CG5102	<i>daughterless</i>	<i>da</i>
FBgn0266724	CG5161	<i>TRAPP subunit 20</i>	<i>Trs20</i>
FBgn0267378	CG7085	<i>sauron</i>	<i>sau</i>
FBgn0267487	CG9181	<i>Protein tyrosine phosphatase 61F</i>	<i>Ptp61F</i>
FBgn0267912	CG9819	<i>Calcineurin A at 14F</i>	<i>CanA-14F</i>
FBgn0086371	CG9829	<i>poly</i>	<i>poly</i>
FBgn0267350	CG10260	<i>Phosphatidylinositol 4-kinase III alpha</i>	<i>PI4KIIIalpha</i>
FBgn0267698	CG10295	<i>p21-activated kinase</i>	<i>Pak</i>
FBgn0283462	CG18279	<i>Immune induced molecule prepropeptide</i>	<i>IMPPP</i>
FBgn0267339	CG33338	<i>p38c MAP kinase</i>	<i>p38c</i>
FBgn0085506	CG40635	-	<i>CG40635</i>

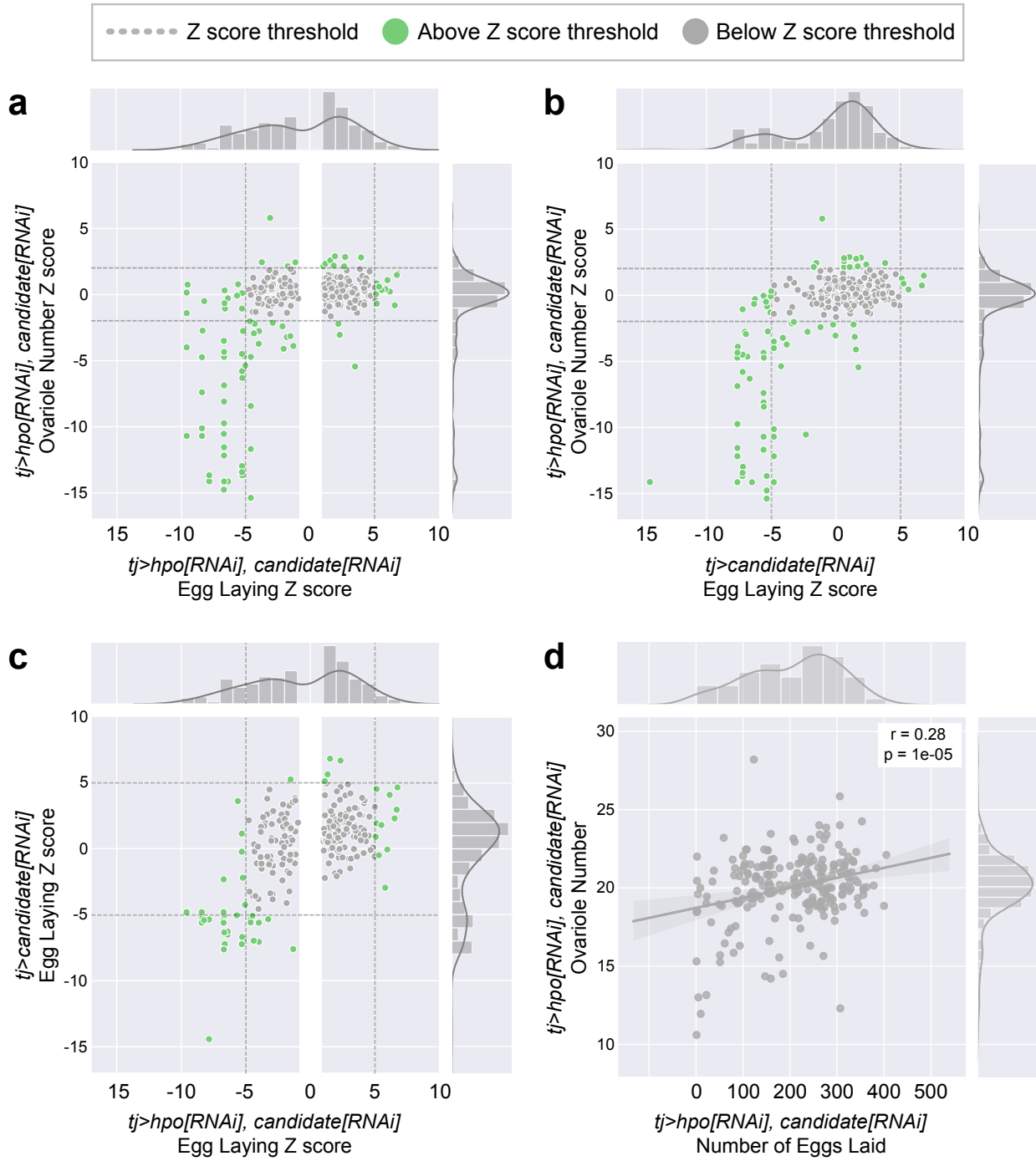
Table 3.3: Candidates with no genetic lines. 12 signalling candidate genes with no available RNAi lines at either BDSC or VDRC at the time of this study.

3.2.2 Ovariole number is weakly correlated with egg laying

Standardization of the results from the three screens using Z scores allowed us to compare the effects of individual genes on one or both of egg laying and ovariole number. We performed a pairwise comparison of the Z_{gene} values for all combinations of screens, and considered genes with $|Z_{gene}|$ values that were above the thresholds set for the phenotype in each screen (above two for ovariole number, above five for egg laying; green dots in Figure 3.2a-c). Across all three screens, loss of function of our positive candidates yielded reductions in ovariole number and egg laying more commonly than increases (Figure 3.2a-c). Comparing the $|Z_{gene}|$ values of egg laying and ovariole number of *tj:Gal4>hpo[RNAi]*, *signalling candidate[RNAi]* adult females revealed that genes that caused a change in egg laying did not always similarly affect ovariole number, and vice versa (Figure 3.2a). We therefore hypothesise that egg laying and ovariole number may be regulated by genetically separable mechanisms. This hypothesis notwithstanding, we observed a weak but statistically significant correlation between egg laying and ovariole number ($p=1e10^{-5}$; Figure 3.2d), and this correlation was most significant in adult females that had a drastic reduction in both phenotypes (Figure 3.2a).

Figure 3.2 (following page): Relationship between Egg Laying and Ovariole Number phenotypes generated in the screens. a) Scatter plots of the Z score for each gene (Z_{gene}) of egg laying versus the ovariole number of adult *tj>hpo[RNAi], candidate[RNAi]* females. **b)** Scatter plots of the Z score for each gene (Z_{gene}) of egg laying of adult *tj>candidate[RNAi]* females versus the ovariole number of adult *tj>hpo[RNAi], candidate[RNAi]* females. **c)** Scatter plots of the Z score for each gene (Z_{gene}) of egg laying of adult *tj>candidate[RNAi]* females versus egg laying of adult *tj>hpo[RNAi], candidate[RNAi]* females. In **a**, **b** and **c**, bar graphs on the top and right sides of each panel show the distribution of genes in each axis of the adjacent scatter plots. Green dots = genes that meet the Z_{gene} threshold for the indicated phenotype. Grey dots = genes that do not meet the Z_{gene} threshold for the indicated phenotype. Dark grey dotted lines = thresholds for each phenotype: $|Z_{gene}| > 5$ for Egg Laying and $|Z_{gene}| > 2$ for Ovariole Number. In **a** and **c**, the white vertical bar removes all genes in the *tj>hpo[RNAi], candidate[RNAi]* with a $|Z_{gene}| < 1$ for egg laying. These genes were not measured in the other two conditions and are therefore not represented in the scatter plots. **d)** Correlation between non-zero Ovariole Number and Egg Laying values.

Figure 3.2: (continued)



3.2.3 No single signalling pathway dominates regulation of ovariole number or egg laying

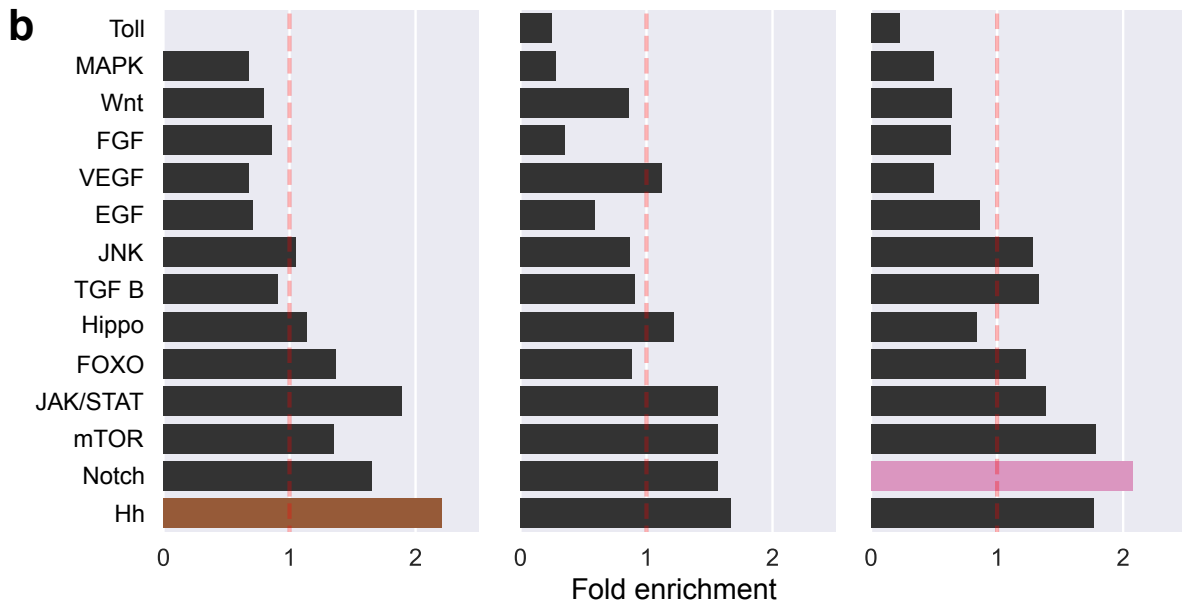
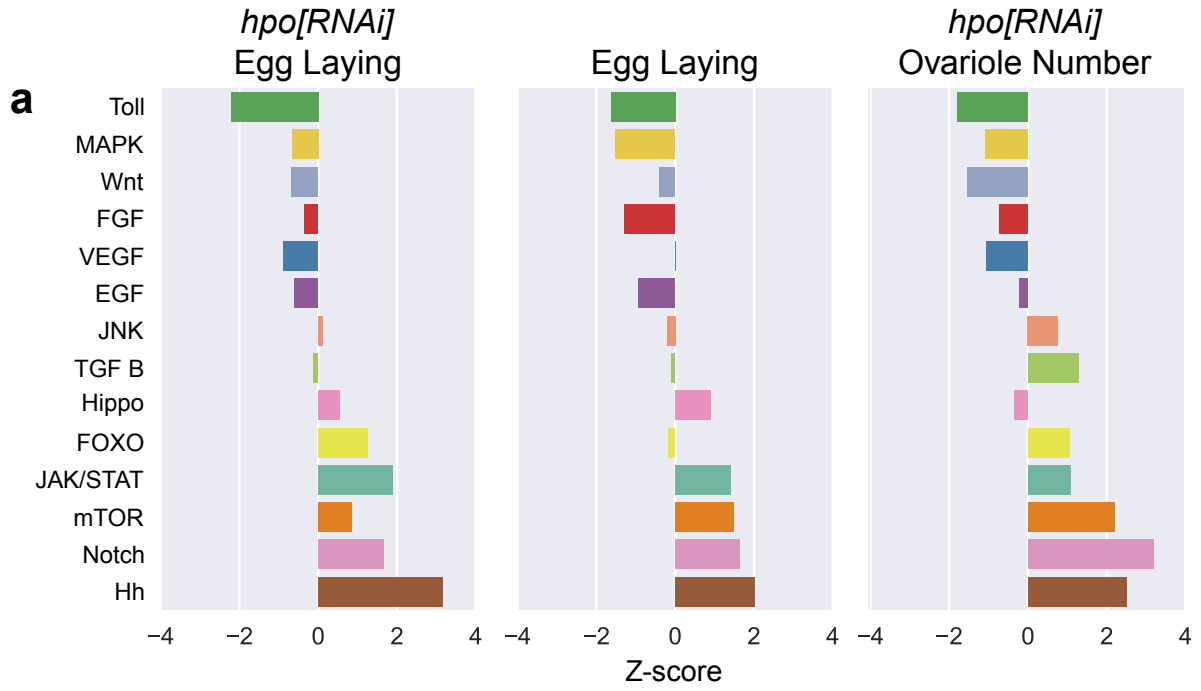
We found that at least some genes from all tested signalling pathways could affect both egg laying and ovariole number (Figure 3.3). To determine if some pathway(s) appeared to play a more important role than others in these processes, we asked whether any of our screens were enriched for genes from a specific signalling pathway. To measure enrichment, we compared the distribution of individual pathway genes among the positive candidates in each screen, to a randomly sampled null distribution of pathway genes among a group of the same number of genes randomly selected from our curated list of 463 signalling genes (Figure 3.3a). Involvement of a pathway in the regulation of a phenotype would be reflected in a difference between the representation of pathway genes in an experimentally derived list and a randomly selected group of signalling genes. We found that rather than only one or a few pathways showing functional evidence of regulating ovariole number or egg laying, nearly all pathways affected both phenotypes (Figure 3.3a). We further tested this result by calculating the hypergeometric p-value for the enrichment of each signalling pathway, in each of the three groups of genes. Consistent with the results of the random sampling approach (Figure 3.3a), we found that most pathway members were not significantly enriched for egg laying or ovariole number phenotypes (Figure 3.3b). The absence of significant enrichment of any specific pathway is not simply attributable to the pool of genes that were screened, because our experimental manipulations of ovariole number and egg laying did cause a change in the distribution of signalling pathway members (Figure C.2). Instead, both phenotypes appeared to be regulated by members of most or all signalling pathways (Figure 3.3). The only two exceptions to this trend were a greater than twofold enrichment of (1) genes from the Notch signalling pathway in the regulation of ovariole number (p-value < 0.05, pink bar in Figure 3.3a, b), and (2) members of the Hedgehog (Hh) signalling pathway in the regulation of Hippo-dependent egg laying (p-value < 0.05, brown bar in Figure 3.3a, 3b; Figure C.3). In summation, our analyses of the enrichment of signalling pathways within the different screens indicated that both ovariole number and egg laying are regulated by genes from nearly all described animal signalling pathways (Figure 3.3a), rather

than being dominated by any single pathway.

Comparing the results of the Egg Laying screens performed in a wild type background (Figure 3.1b) or in a *hpo[RNAi]* background (Figure 3.1a), revealed that most of the genes that met a threshold of $|Z_{gene}| > 5$ in one screen, did not meet that threshold in the other screen (Figure 3.2c). This result suggests the existence of both Hippo-dependent and Hippo-independent mechanisms of regulation of egg laying. The interpretations of separable Hippo-dependent and -independent regulation of egg laying, and of the separable regulation of ovariole number and egg laying, was supported by the results of the network analysis described in the following section.

Figure 3.3 (following page): Enrichment of genes of individual signalling pathways among the experimentally obtained positive candidates of each screen. a) Enrichment/depletion analysis to identify over- or under-represented members of individual signalling pathways among positive candidates of each screen. Positive Z scores represent an enrichment, and negative Z scores represent depletion, of genes of a pathway among those genes that experimentally affected the phenotype. Enrichment and depletion are defined relative to a null distribution of the expected number of members of a signalling pathway among a group containing the same number of randomly selected signalling genes. **b)** Fold enrichment and hypergeometric p-value calculation to identify over- or under-representation of the genes of a pathway in each screen. Significantly enriched pathways (coloured bars: brown = Hedgehog; pink = Notch) are defined by having a hypergeometric p-value less than 0.05. Enrichment/depletion analysis of the 273 signalling pathway genes above the threshold $|Z_{gene}| > 1$ (Figure 3.1a) before screening is illustrated in **Figure C.2**. **Figure C.3** compares the Z_{gene} of egg laying of adult females of *tj>hpo[RNAi],candidate[RNAi]* plotted against Z_{gene} of egg laying of *tj>candidate[RNAi]* adult females displayed by pathway.

Figure 3.3: (continued)



3.2.4 Centrality of genes in the ovarian protein-protein interaction networks can predict the likelihood of loss of function phenotypic effects

The finding that these reproductive traits were regulated by the genes of all signalling pathways led us to consider the broader topology of putative gene regulatory networks in the analysis of our data. Previously characterized genes in the ovary are often pleiotropic and can regulate both ovariole number and egg laying^{26,30}. As with proteins in a linear pathway, proteins in a protein-protein interaction network (PIN) are more likely to function in conjunction with genes that are connected to them within the network (e.g. Ideker and Sharan⁵⁰, Jeong et al.⁵¹). Centrality is one measure of the connectedness of a gene in the PIN and can be used to identify the most important functional centres within a protein network^{52,53}. Most centrality measures use path length, which is a measure of the number of other proteins required to link any two proteins in the network. Here we used four commonly used metrics to quantify gene centrality, each measuring slightly different properties^{54,55}. (1) *Degree centrality* is proportional to the number of proteins that a given protein directly interacts with. (2) *Betweenness centrality* measures the number of shortest paths amongst all the shortest paths between all pairs of proteins that require passing through a particular protein. (3) *Closeness centrality* measures the average shortest path that connects a given protein to all other proteins in the network. (4) *Eigenvector centrality* is a measure of the closeness of a given protein to other highly connected proteins within the network.

We hypothesised that if the candidate genes we identified in our screen as playing roles in ovarian function worked together as a PIN, then the degree of centrality of a gene might be an indicator of function. To test this hypothesis, we first compiled a PIN consisting of all described interactions between *D. melanogaster* proteins, from the combination of publicly available protein-protein interaction (PPI) studies in the DroID database (see Methods). We then calculated the four centrality measures described above for all genes within the *D. melanogaster* PIN (Table C.1). We rank ordered only the genes tested in each screen by their score for each centrality measure, and asked whether their rank order correlated with the results of the screen,

plotting these results as a receiver operating characteristic (ROC) curve. Positive correlations between centrality (a continuous variable) and phenotype (a binary variable: above or below the $|Z_{gene}|$ threshold) are reflected in an area under the curve (AUC) of more than 0.5. We found that the higher the centrality score, the greater the likelihood that a gene had $|Z_{gene}|$ values above our threshold for effects on ovariole number and egg laying (Figure 3.4a; Table C.2). This supports the premise that the positive candidates identified in our screen function together as a network in the regulation of either ovariole number or egg laying. Interestingly, while the centrality of genes did predict whether a gene would affect our phenotypes of interest, it could only weakly predict the strength of that effect (p-value < 0.05 in Figure C.4).

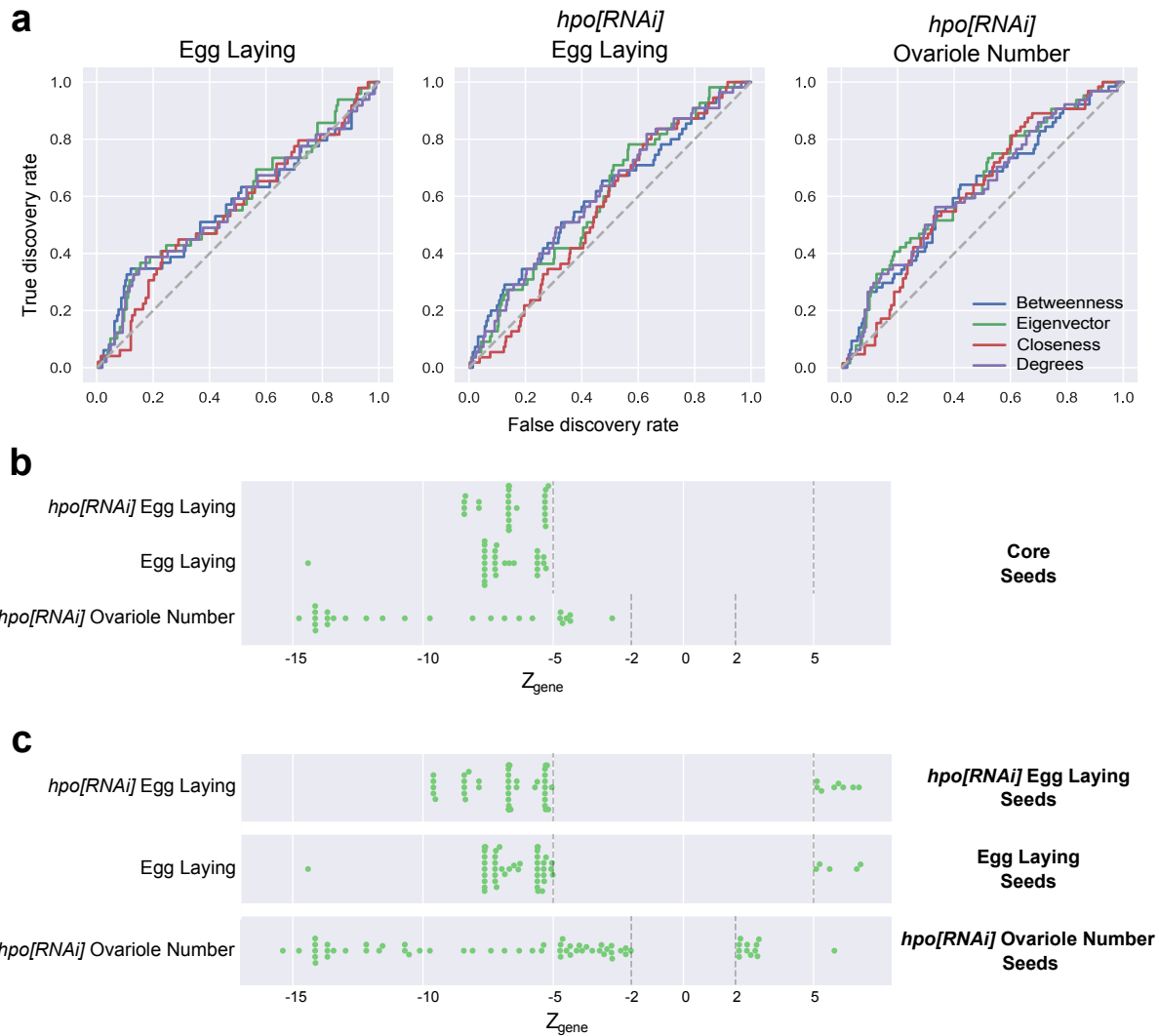


Figure 3.4: Screened genes function as a network. a) Receiver operating characteristic (ROC) curves of genes ordered by rank for each of four network centrality metrics (Betweenness centrality, Eigenvector centrality, Closeness centrality and Degree centrality) versus a binary outcome (above or below Z score threshold) for each of the three screens. For each screen and metric, the Area Under the Curve (AUC) is > 0.5 (Table C.2). **b)** Genes whose $|Z_{gene}|$ value was above the threshold (green dots; Table 3.2) in all three screens were assigned to the Core seed list. **c)** Genes whose $|Z_{gene}|$ value was above the threshold (green dots; Table 3.2) in each screen were assigned to the corresponding seed list. **Table C.2** Tabulates the AUC values for the ROC curves for each centrality measure for the three screens (Figure 3.4a). **Figure C.4** Compares the distribution of Z_{gene} scores of the positive candidate genes for the first and fifth quintile genes sorted by their centrality metrics. **Figure C.5** compares the network metrics of the four gene lists in Figure 3.4b to two null distribution of genes selected from the PIN.

3.2.5 Genes regulating egg laying and ovariole number regulation form non-random gene interaction networks

The centrality analyses above suggested that the genes implicated in ovariole number and egg-laying displayed characteristics of a functional network. PINs can often be further sorted into a collection of sub-networks. A sub-network is a smaller selection of proteins from the PIN. Examples of such sub-networks could be proteins within the same subcellular organelle⁵⁶ or genes that are expressed at the same time⁵⁷, thus making them likely to function together⁵⁸. A putative module is a sub-network that can perform regulatory functions as a unit, independent of other sub-networks, and has key measurable features^{44,59,60,61}. Genes and interactions between genes are not mutually exclusive to such putative modules and can be shared between putative modules. We therefore asked if our sub-networks, consisting of genes that showed similar mutant phenotypes, might display features of modularity. To determine whether genes that were implicated in regulation of ovariole number and egg laying interacted with each other in specific groups more than would be expected by chance, we created four lists of genes, called "seed" lists, based on their individual phenotypic effects based on our screen results: (1) the core seed list, including genes positive in all three screens (Figure 3.4b); (2) the egg laying seed list, including genes positive in the wild type background egg-laying screen (Figure 3.1b; Figure 3.4c); (3) the *hpo*[RNAi] egg laying seed list, including genes positive in the *hpo*[RNAi] background egg laying screen (Figure 3.1a; Figure 3.4c); and (4) the *hpo*[RNAi] ovariole seed list, including genes positive in the *hpo*[RNAi] background ovariole number screen (Figure 3.1c; Figure 3.4c). Interestingly, the core seed list, comprising genes that affected all three measured phenotypes, only consisted of genes that caused a reduction in both ovariole number and egg laying (Figure 3.4b).

We then asked whether these four seed lists were more connected than would be expected by chance. In other words, we formally tested them for modularity as defined above. Meeting our criteria for modularity would suggest that the genes in these phenotypically separated seed lists might operate together as putative functional modules within the *Drosophila* PIN. We performed our modularity test using four commonly measured network metrics: (1) Largest Connected

Component (LCC) (the number of proteins or nodes connected together by at least one interaction), (2) network density (the relative number of edges as compared to the theoretical maximum), (3) total number of edges, and (4) average shortest path (average of the minimum distances connecting any two proteins). We considered a sub-network to show modular features if they showed most of the following properties: higher LCC, higher network density, more edges, and shorter average shortest path length when compared to a similarly sized, randomly sampled selection of genes from the PIN.

To determine whether these criteria would correctly identify signalling genes, which are known to function together as a module, we measured these four parameters in the original set of genes (all signalling genes) used in this study (Table 3.1). We found that the signalling genes display features of modularity when compared both to a randomly selected set of genes, as well as to a degree-controlled list of genes (Figure C.5a). We then used this approach to test the modularity of the four phenotypic sub-networks, when compared to two different "control sub-networks" consisting of a group of the same number of genes as contained in the sub-network, one chosen randomly from among the candidate genes from our initial screen list (Table 3.1), and the second chosen from a degree-controlled list of genes selected from the entire PIN (see Methods: Building degree-controlled randomized networks). We found that the four predicted phenotypic sub-networks showed higher LCC, higher network density, more edges (Figure C.5b), compared to both "control sub-networks". This result suggests that these sub-networks display many features of modularity (although their average shortest path length is higher than controls, rather than lower) and may function as putative modules within the PIN to regulate one or both of ovariole number or egg laying.

Based on published molecular interactions, in addition to the four criteria described above, further evidence for putative functional modules of genes can also be obtained by applying algorithms that use either the shortest path method⁶² or the Steiner Tree approach⁶³. Such methods identify and predict functional connections between the seed proteins, as well as additional nodes (proteins or genes) that have not been experimentally tested within the given parameters, but are known to interact with the seed genes in the PIN^{64,65}. This process can provide evidence for or against the existence of a predicted functional module, and subsequent

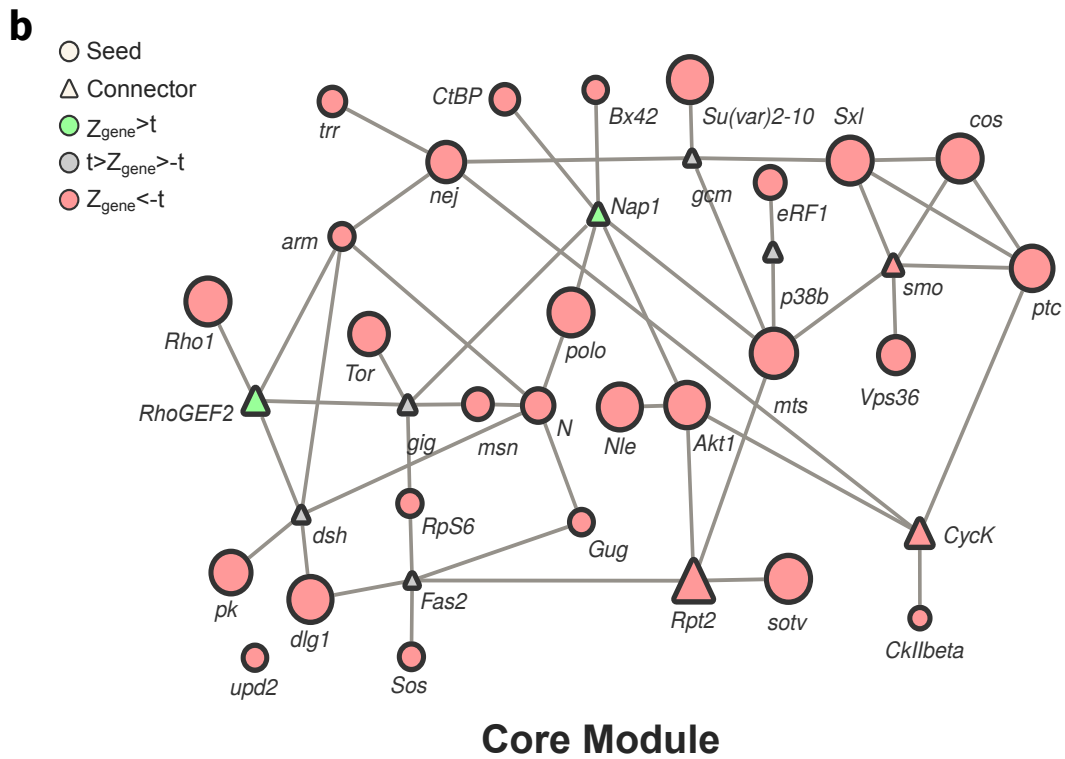
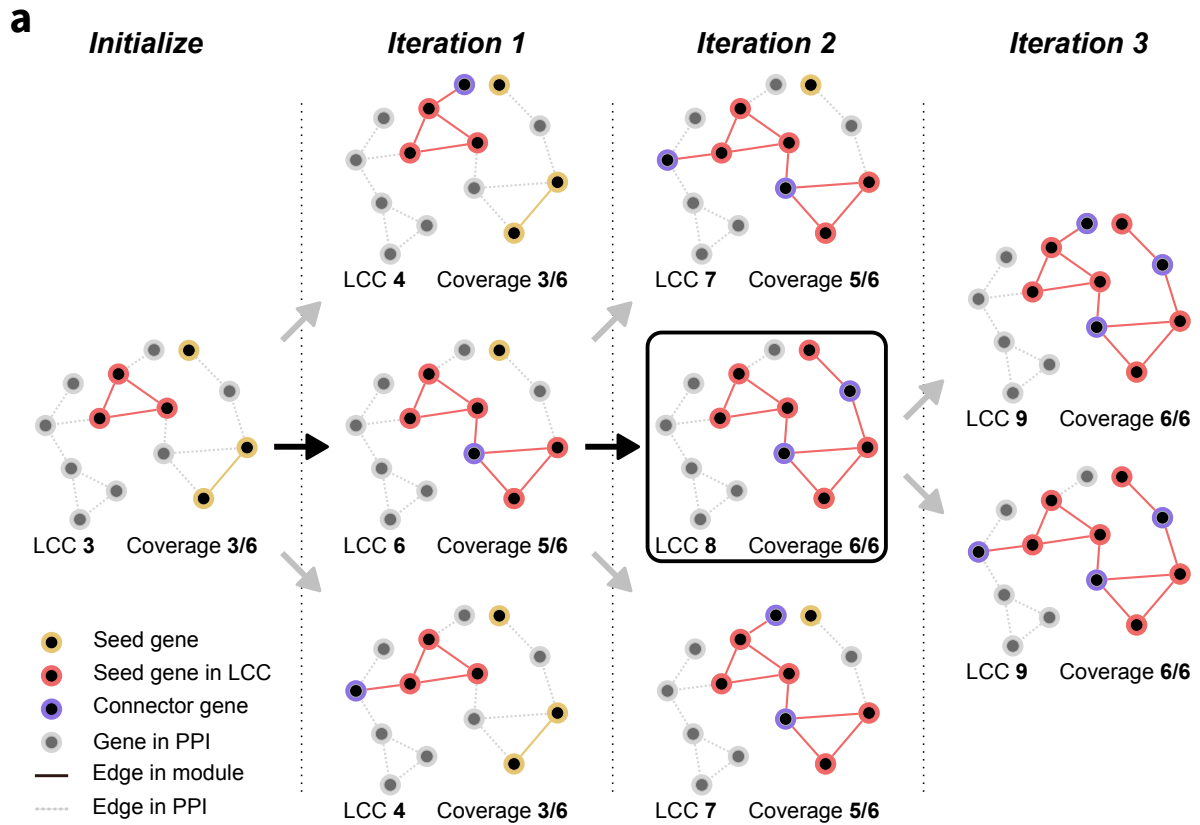
experimental testing of this predicted module can confirm its functionality. Given its recent success in predicting gene modules, we applied the previously published Seed Connector Algorithm (SCA), a member of the Steiner Tree algorithm family^{66,67}, to the groups of genes that had similar phenotypic effects in our screens (seed genes; Figure 3.4b, c). The SCA connects seed genes and previously untested novel genes (connectors) to each other using a known PIN, producing the largest possible connected putative module given the data. Using the PIN and the aforementioned four lists of seed genes, we applied a custom python implementation of the SCA (Methods: 04_Seed-Connector.ipynb) to build and extract the largest possible (given our PIN) connected putative modules that regulate egg laying and ovariole number.

This SCA method yielded four putative modules, one for each seed list, which we initially referred to as the Core Module (Figure 3.5b), *hpo[RNAi]* Egg Laying module (Figure C.6), Egg Laying Module (Figure C.7), and *hpo[RNAi]* Ovariole Number Module (Figure C.8) respectively. Each of these four putative modules contained seed genes, which had been functionally evaluated in our screens (green and red circles in Figure 3.5), as well as connector genes, which were genes newly predicted as regulators of these phenotypes (green and red triangles in Figure 3.5). Of the four putative modules generated by the SCA, we found that the Core module had higher centrality measures than the other three putative modules (Figure C.9). We interpret this to mean that the genes regulating these "Core phenotypes" are more strongly connected to each other.

We found, however, that these four groups of genes produced by the SCA did not have increased LCC values, increased network density, more edges nor decreased average shortest path (Figure C.10), compared to our "control sub-networks". This result shows that the SCA in this instance does not provide evidence for putative functional modules from the four phenotypic sub-networks in our system, above and beyond the evidence provided by the application of the four network metrics discussed above. To be conservative in our description of these results, we therefore henceforth refer to these four groups of genes united by phenotype and with strong predicted interactions, as sub-networks rather than as modules. We noted that each of these four sub-networks contains genes from most, if not all, known signalling pathways, rather than only genes from a single pathway (Figure C.11).

Figure 3.5 (following page): Representation of Seed Connector Algorithm and output. a) Schematic representation of the seed connector algorithm. The algorithm initializes by creating a sub-network of seed genes from the PIN and computes the Largest Connected Component (LCC) and coverage (number of genes from the seed set in the LCC). At each iteration, genes in the direct neighbourhood of the LCC (distance = 1) are added one at a time to the seed set, and the coverage and LCC are re-computed. This process is repeated for each gene in the direct neighbourhood, each time restarting from the seed set of the preceding iteration. If any gene outside the seed set but in the direct neighbourhood is found to maximize coverage while minimizing the LCC, it is added to the seed set as a connector gene. Black arrows indicate the path taken by the algorithm for which the criteria of maximal coverage and minimal LCC are met; such a path would be used to proceed to the subsequent iteration. Grey arrows indicate paths that fail to meet these criteria; such paths would be disregarded. The iteration repeats until the coverage cannot be increased; in this schematic example, this state is achieved in iteration 3. **b)** The Core sub-network generated by the Seed Connector Algorithm (SCA) based on the results of the genetic screens (Figure 3.1a-c). The size and colour of the shapes indicate the relative Z_{gene} score of ovariole number of adult *tj>hpo[RNAi]*, *candidate[RNAi]* females. Circles indicate seed genes (functionally tested in the screen; Table 3.2; Table C.1) while triangles are connector genes (novel predicted genes; Table C.1, Figures C.6 to C.8). Green = genes with a positive Z_{gene} score above the threshold; red = genes with a negative Z_{gene} score above the threshold; grey = genes with Z_{gene} values below the threshold. *hpo[RNAi]* Egg Laying, Egg laying and *hpo[RNAi]* Ovariole Number sub-networks generated by the Seed Connector Algorithm (SCA) are illustrated in **Figures C.6 to C.8** respectively. **Figure C.9** shows the distribution of the four centrality measures calculated for the genes in each of the four phenotypic sub-networks obtained from the SCA (Figure 3.5 and Figures C.6 to C.8). **Figure C.10** Compares the network metrics after application of the SCA. **Figure C.11** shows the enrichment/depletion of signalling pathway genes in each of the four sub-networks obtained from the SCA.

Figure 3.5: (continued)



3.2.6 Low edge densities between sub-networks suggest genetically separable mechanisms of ovariole number and egg laying

Our network analysis identified four highly connected sub-networks of genes that regulate two distinct developmental processes, together with or independently of Hippo signalling activity: ovariole number determination, which occurs primarily during larval development, and egg laying, which takes place in adult life (Figure 3.5). We wished to assess the degree to which there might be any shared genetic components between these four phenotypic sub-networks, and whether the addition of connector genes by the SCA had any impact on this. To understand potential interactions between the phenotypic sub-networks in the regulation of both ovariole number and egg laying, we constructed a composite network of all genes in each of the four phenotypic sub-networks (Figure 3.5b; Figures C.6 to C.8), which we refer to as the "meta network" (Figure 3.6a). We then grouped the genes of the meta network into seven bins based on their phenotypic effects as measured in the three screens, resulting in sub-groups I through VII shown in Figure 3.6a. To ask whether the genes in these phenotypic groupings showed any notable interaction patterns, we compared the connectivity between genes assigned to the same phenotypic group, to the connectivity of a group of the same size randomly assembled from the genes of the meta network (Figure C.12). As a measure of connectivity, we used an edge density map, which reflects the number of interactions between the genes within a group and between groups. We quantified the deviation between the edge density of each of groups I through VII, and their corresponding randomly assigned groups of the same size, by computing their respective Z score. When we included only seed genes in each of groups I through VII, we found that the edge density values of these groups were somewhat lower (Z score < -1.5) than those of the randomized groups (Figure C.12a). The single exception to this was group IV, whose members shared more edges with each other than did the members of its randomized comparison group of the same size (Figure C.12a). In other words, these groups of phenotypically binned seed genes were not notably more connected to each other than we would expect by chance.

In contrast, expanding each of the seven phenotypic sub-groups to include both seed genes and

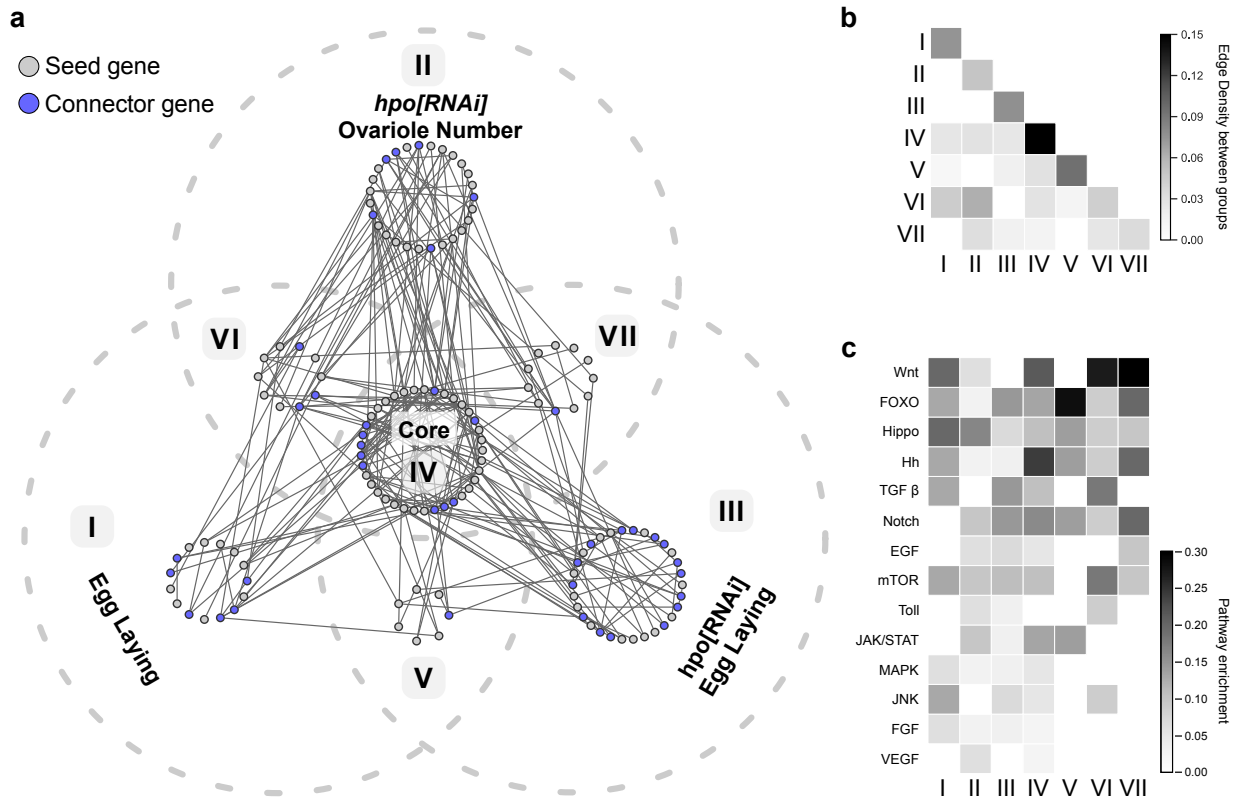


Figure 3.6: Phenotypically separable sub-networks formed by analysis of the combined genes from all sub-networks. The meta network is generated by the union of the genes in the four phenotypic sub-networks: *hpo[RNAi]* Egg Laying (Figure C.6), Egg Laying (Figure C.7), *hpo[RNAi]* Ovariole Number (Figure C.8) and Core (Figure 3.5b). **a**) The meta network is represented as a Venn diagram, in which each grey dotted outline represents the screen in which a given gene was identified as affecting the scored phenotype. Within each sub-network, grey circles indicate seed genes, and blue circles indicate connector genes. Solid grey lines indicate interactions between genes in the meta network from the PIN. **b**) Edge densities between the seven sub-networks of the meta network. **c**) Relative enrichment of screened members of the 14 tested developmental signalling pathways within the seven sub-networks of the meta network. **Figure C.12** compares the edge density of the seven sub-networks of the meta network to the edge densities of a random assignment of the positive candidates in the screen to a seven similarly sized sub-networks.

the connector genes predicted by the SCA changed the edge densities of these groups relative to their randomized control groups. Specifically, edge densities were much lower between groups I, II and III (Z score < -3), and much higher within group IV (Z score > 3) (Figure C.12b). This shows that applying the SCA to these phenotypically binned groups increased the non-random differences in connectivity between them that were already present within the seed genes (Figure C.12a), thus clarifying the internal structure of the meta network.

We then asked if these seven sub-groups were as connected to each other, as were the genes within each of the sub-groups, again using the edge density assessment as described above (Figure 3.6b). This analysis yielded three principal findings. First, edge densities between the three groups corresponding to the three scored phenotypes (I, II and III in Figure 3.6a) were very low (Figure 3.6b). This implies that the genes in each of the groups that regulate only one phenotype (I, II and III in Figure 3.6a) share more interactions with themselves than with genes in the other two groups, suggesting that each of these initially scored phenotype can be largely regulated by an independent, non-interacting set of genes. Second, the core group (IV in Figure 3.6a) displayed a higher edge density with the other three groups (I, II and III in Figure 3.6a) than any of those three groups did with each other (Figure 3.6b). Consistent with the definition of core genes as regulating all three scored reproductive phenotypes, this result suggests that the core genes, in contrast to those from the other three groups, may share substantial functional interactions with genes of the other groups. Finally, three small additional groups emerged from this analysis (V, VI and VII in Figure 3.6a), suggesting small networks of genes that might work together to regulate two of the three scored phenotypes. In sum, this meta network analysis supports the hypothesis of three potentially largely non-interacting genetic networks that regulate Hippo-dependent ovariole number, Hippo-dependent egg laying, and Hippo-independent egg laying respectively. The presence of smaller sub-networks (V, VI and VII in Figure 3.6a) that interact with each other further supports the observation that the putative modules predicted by the SCA – which we refer to as sub-networks – could include genes that function within more than one such sub-network (Figure C.9). Moreover, each of these genetically separable sub-networks included genes in multiple signalling pathways (Figure 3.6c).

3.2.7 Network analysis predicts novel genes involved in egg laying and ovariole number

The four predicted phenotypic sub-networks produced by the SCA approach included connector genes that were not included in our original screen, and thus had not been tested for possible effects on our phenotypes of interest (triangles in Figure 3.5b; Figures C.6 to C.8). Given that prior work in human disease models showed that predicted disease modules can correctly predict gene involvement in the relevant diseases^{66,67,68,69}, we asked whether our deployment of the SCA had likewise successfully p:*RNAi* lines for each connector, driven by *tj:Gal4* to measure the effects of knocking down each of the connector genes (triangles in Figure 3.5b and Figures C.6 to C.8) both on phenotypes within the sub-network where they were predicted (Figure 3.7a-b, Table C.3), and on either of the other two tested phenotypes (Figure 3.7c, Table C.4).

Of the ten predicted novel connectors within the Core sub-network, loss of function of several of these had significant effects on at least one of the three scored phenotypes. Five affected ovariole number, two affected Hpo-dependent egg laying and one affected Hpo-independent egg laying. However, only one of them significantly altered all three scored phenotypes (Figure 3.7a; Table C.3).

The predicted connector genes from two of the other three phenotypic sub-networks showed high positive prediction rates for novel genes within the sub-networks. RNAi against seven out of 18 of the *hpo[RNAi]* Egg Laying connectors, three out of 11 of the *hpo[RNAi]* Ovariole Number connectors, and none of the 11 Egg Laying connectors, significantly affected the sub-network phenotype (Table C.3). Thus, although the Egg Laying connectors failed to impact this phenotype in our assay, 41.1% and 27.2% of the connectors from the other two sub-networks were correctly predicted (Figure 3.7b; Table C.3).

In sum, taken across all sub-networks, this methodology correctly identified genes regulating at least one of the scored reproductive phenotypes, at significantly higher rates than those obtained in the original screen of 463 members of all known signalling pathways (Figure 3.7c; Table C.4). By this measure, testing network-predicted novel genes derived from experimentally

obtained data was even more successful than testing signalling pathways as a means of identifying novel genes that regulate ovariole number and egg laying.

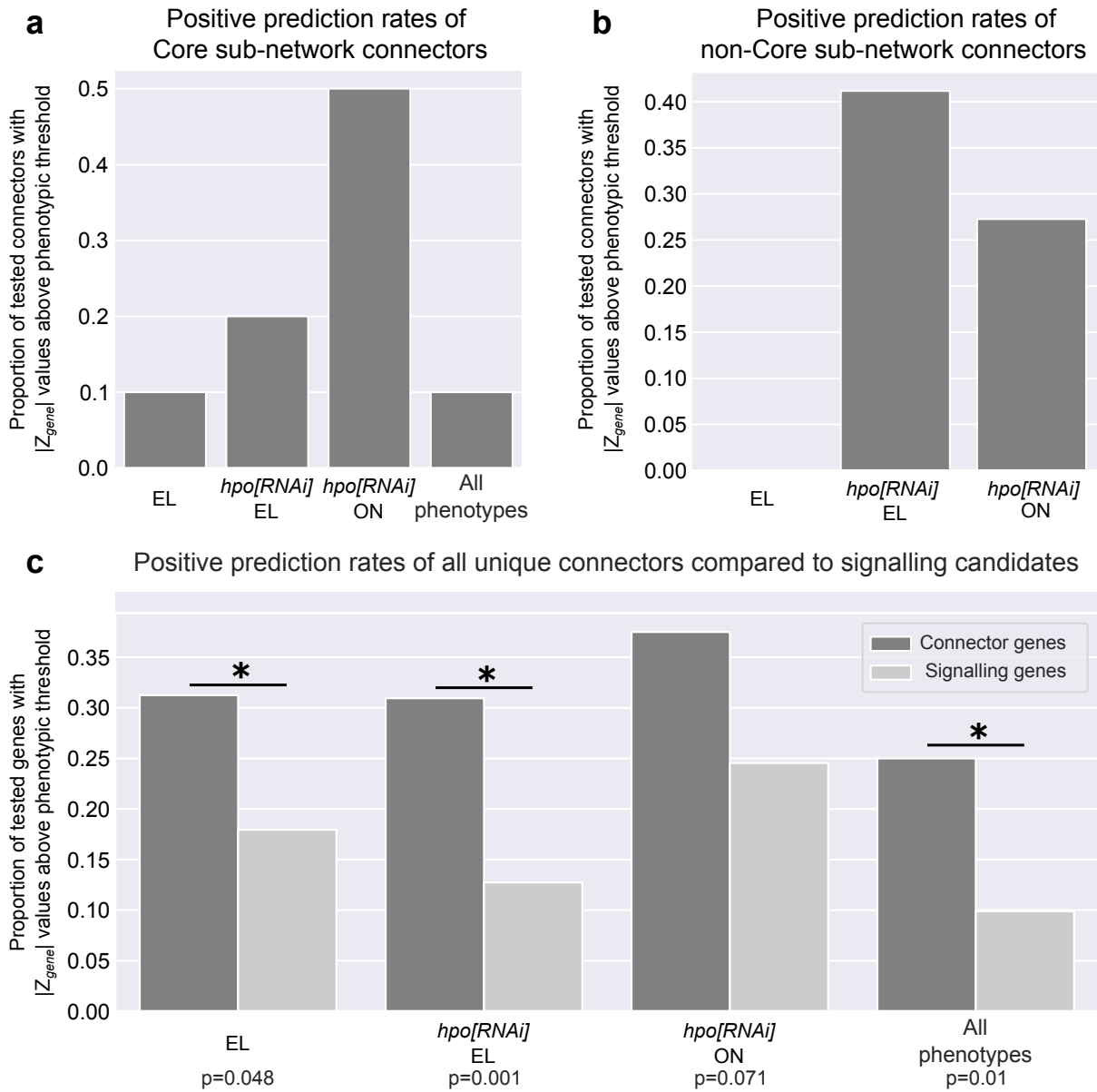


Figure 3.7: Positive prediction rates of the connector genes in each of the four sub-networks. **a)** Proportion of Core sub-network connector genes with $|Z_{gene}|$ above the threshold in each of the three screens. The "All phenotypes" category includes the genes with $|Z_{gene}|$ above the threshold in all three screens. **b)** Proportion of tested connector genes in each of the three sub-networks with $|Z_{gene}|$ above the threshold within their corresponding screen. **c)** Proportion of all unique connector genes (dark grey bars) predicted by all four sub-networks compared to the proportion of signalling candidate genes (light grey bars) with $|Z_{gene}|$ above the respective threshold in any of the three phenotypic screens (Figure 3.1a-c). Positive connector and signalling candidates that were above the $|Z_{gene}|$ threshold in all three phenotypic screens (Figure 3.1a-c) are indicated in an "All phenotypes" column. Statistical significance was computed using the binomial test, comparing the probability of a positive candidate amongst the connectors to the probability of a positive candidate amongst the signalling candidates (p-value is found below each bar). **Table C.3** tabulates the distribution of seed genes and connectors in each sub-network used in Figure 3.7a, b. **Table C.4** tabulates the unique connector and signalling genes above $|Z_{gene}|$ threshold for the three phenotypic measurements plotted in Figure 3.7c. The connector genes are listed in the Table C.1 and can be identified under the column header **[SubNetworkName]_Connector**. The raw data for egg laying and ovariole number for each of the connector genes can be found within the Table C.1.

3.3 Discussion

In this study, we have identified many novel genes that regulate either or both of egg laying and ovariole number. Though the development of the insect ovary has been studied for over 100 years, our understanding of the genetic mechanisms that regulate the development of the ovary is sparse. The female reproductive system and its ability to produce eggs are one of the key determinants for the survival of a species in an ecological niche. The genes we have uncovered here are possible targets for the regulation of the construction and function of the reproductive system in *Drosophila melanogaster*, and potentially in other species of insects as well.

Understanding the gene regulatory networks that regulate egg laying and ovariole development could provide a framework to understand the key regulatory steps during this process that may be modified over evolutionary time, to yield the wide diversity of ovariole numbers and fecundities displayed by extant insects. We suggest that, given our success in applying a network approach to the results of a traditional forward genetic screen, the field of developmental genetics should find it fruitful to apply network analyses to the interpretation of large scale transcriptomic and proteomic data.

3.3.1 Identification of regulatory sub-networks for ovariole development and egg laying

The *D. melanogaster* ovary is a commonly studied model for organogenesis^{15,21,25}, stem cell maintenance³⁰ and interactions of development and ecology^{5,12,70,71}. Nevertheless, our understanding of the genetic mechanisms that regulate these processes remains fragmentary. In this paper, we have identified four distinct protein interaction sub-networks that regulate ovariole number and egg laying in the *D. melanogaster* ovary. These sub-networks consist of both novel and previously characterized genes that regulate either ovariole number or egg laying or both, thus enhancing our understanding of the genetic underpinnings of this reproductive system.

Of the four sub-networks, the Core sub-network affects both ovariole number and egg laying. The Core sub-network contains numerous housekeeping genes, including regulators of

transcription, translation and cell division such as *polo*⁷², *cyclin K*⁷³, *nucleosome assembly protein 1*⁷⁴ and *eukaryotic translation release factor 1*⁷⁵. While *polo* and *eukaryotic translation release factor 1* are members of signalling pathways, *cyclin K* and *nucleosome assembly protein 1* are genes predicted by the SCA. Given that the Core sub-network largely consists of genes whose loss of function decreases both of these parameters, we hypothesise that these are essential genes for the basic structure and function of the ovaries. Essential genes are more interconnected in a PIN with higher centrality measures (⁵¹; but see ⁷⁶) and interestingly, we find that the genes in the Core sub-network also have higher connectivity than those in the other three sub-networks (Figure C.9).

In addition to genes that regulate basic cellular processes, the Core sub-network is enriched for the central components of the Hh signalling cascade, namely *patched (ptc)*, *smoothened (smo)* and *costa (cos)*⁷⁷. However, we find that the loss of Hh ligand, which is expressed in the TF cells in the developing larval ovary⁷⁸, does not significantly affect either ovariole number or egg laying. Though surprising, ligand-independent activation of Hedgehog signalling has been observed before. For example, in the *Drosophila* eye, loss of either *ptc* or *cos* in clones leads to non-cell autonomous proliferation in wild type cells, as well as growth disadvantages in the mutant tissue⁷⁹. In another example, sufficient intracellular *smo* levels can also activate downstream transcription of Hh pathway targets, showing that Hh itself is not always required to activate the cascade⁸⁰.

3.3.2 Development of the larval ovary

The *hpo*[RNAi] Ovariole Number sub-network is composed of genes that affect the Hippo signalling activity-dependent determination of ovariole number during development. Establishment of ovariole number occurs largely during the third instar stage of larval development in *D. melanogaster*^{20,21,23,38}. During this period, the TFCs are specified in the anterior of the ovary and undergo rearrangement into stacks of cells called TFs, each of which gives rise to an ovariole^{21,24}. TF specification requires the expression of *engrailed (En)*⁸¹ and the transcription factors Bab1 and Bab2, encoded by the *bric-à-brac* locus^{82,83}. A third transcription factor, Lmx1a, was recently found to be necessary for the specification of the TFCs⁸⁴. Our

hpo[RNAi] Ovariole Number sub-network identifies numerous additional novel transcription factors including *bunched (bun)* and *retinoblastoma-family protein (rbf)*, which we hypothesize could also be involved in the specification of ovariole number. *bun* and *rbf* have been implicated in the migration⁸⁵ and endoreplication⁸⁶ of the follicle cells during oogenesis, but have not, to our knowledge, been previously identified as playing a role in the context of larval ovary development.

The TFCs specified in the larval ovary undergo a process of convergent extension to form TFs. This process of convergent extension requires cell intercalation, and the actin depolymerizing factor Cofilin, encoded by the gene *twinstar*, is essential to this process²⁵. During intercalation, the cells also dynamically modify their actin cytoskeleton and their expression of E-cadherin²¹. Our *hpo[RNAi]* Ovariole Number sub-network further identifies *Rho1*⁸⁷ and *Rho kinase (Rok)*⁸⁸ as necessary for correct ovariole number. During the extension of the *D. melanogaster* embryonic germ band, a commonly studied model of convergent extension, the localised activation of the actin-myosin network facilitated by *Rho1* and *Rok* is necessary for cell intercalation⁸⁹. Given the known roles of *Rho1* and *Rok* as regulators of the actin cytoskeleton⁹⁰, we propose that TF assembly in the ovary requires both these proteins for correct cell intercalation. A third actin cytoskeleton regulator, *misshapen (msn)*, was also identified by our *hpo[RNAi]* Ovariole Number sub-network. *msn* encodes a MAP kinase previously shown to regulate the polarisation of the actin cytoskeleton during oogenesis⁹¹, but has not, to our knowledge, been studied to date in the context of larval ovarian development.

We propose that the polarity of the somatic cells in the ovary is also necessary for correct larval ovary development, given the presence of the lateral membrane proteins *discs large 1 (dlg1)* and *prickle (pk)* in the ovariole sub-network. During the maturation of the TFs during larval development, the TFCs undergo significant cell shape changes, coincident with localised expression of beta-Catenin and actin to the lateral edges of the TFCs²¹. Restriction of the E-cadherin domain in epithelia requires establishment of the basolateral domain⁹² and we propose that testing a similar requirement for *dlg1* and *pk* in the larval ovary would be a fruitful avenue for future studies.

3.3.3 Network analysis as a tool in developmental biology

Using a systems biology approach to analyse RNAi screening data has proven fruitful, providing us with new insights into the development and function of the *D. melanogaster* ovary by identifying novel and previously understudied genes that regulate this process. Systematic analysis of the function of single genes in development has been a historical convention and has provided valuable and precise genetic interaction information^{93,94}. With the advent of genome-wide analysis, however, we can use data from a larger number of genes to predict the identity of additional functionally significant genes with relative ease⁶⁵. We note that the novel gene prediction rate within each individual sub-network ranged from as high as 41.1% from the *hpo[RNAi]* Ovariole Number sub-network to as low as 0% from the Egg Laying sub-network (Figure 3.7b; Table C.3). We suggest that this may be due to multiple factors. Firstly, the possible incompleteness of the PIN is expected to lead to some areas of the network being sparse or non-existent⁹⁴. If the sub-network of interest happens to fall in such regions of the PIN, prediction algorithms will fail. Secondly, the initial restriction of tested genes to signalling pathway members might have provided a seed list too sparse to usefully predict functional connectors. Finally, it could be the case that "Egg Laying" is such a complex phenotype that its gene regulation cannot be adequately captured within a highly connected network of the type suited for identification by the analyses we have used here.

Ovariole number in *D. melanogaster* is the outcome of a discrete developmental process with a clear beginning and end, comprising a specific series of cellular behaviours that take place in the confines of one organ^{21,23,71}. Once established during larval life, ovariole number in *Drosophila* remains unaltered through to and during adulthood, even if oogenesis within those ovarioles suffers congenital or age-related defects³. Because previous work suggested that ovariole number in *Drosophila* could have at least some predictive relation to egg laying^{6,70}, we reasoned that scoring the latter phenotype in a primary screen (Figure 3.1a) could be an effective way to uncover ovariole number regulators (Figure 3.1c). While our results showed that this was true in many cases, it was also clear that these two traits can vary independently (Figure 3.2), highlighting the fact that ovariole number is not the only determinant of egg laying. Egg-laying dynamics, even during the limited five-day assay used in our study, are likely

influenced not just by a single anatomical parameter such as ovariole number, but rather by many biological, biomechanical, hormonal and behavioural processes. Consequently, the sub-network we were able to extract from the results of this screen (Figure 3.1a, b) might be too coarse to extract novel genes that participate in potentially complex gene interactions regulating egg laying. Furthermore, genes predicted within each of the sub-networks are unlikely to function exclusively within just one sub-network. This conclusion is supported by our observation that genes predicted to function in any of the sub-networks, also function in at least one of the four sub-networks at a higher rate than genes selected for screening by their presence in a signalling pathway (Figure 3.7c). We also observe that though substantial regions of the meta network do not share interactions with genes in the other sub-networks (Figure 3.6b), we do find smaller sub-networks where there is some overlap, further indicating pleiotropy of some genes in both egg laying and ovariole number regulation.

The predictive rates of the approach we have used here, although encouraging, are likely limited by the degree of noise in the high-throughput data used to generate the PIN⁹⁵, the sparseness of the PIN, and the degree of misidentification of protein interactions⁹⁶. Addressing one or more of these parameters could improve the outcomes of future network predictions from developmental genetics data. For example, the problem of sparseness, which is a paucity of high confidence detectable interactions relative to all biologically relevant interactions, has been addressed in other studies by using an "Interolog PIN"⁹⁷ in place of an organism-specific PIN. An Interolog PIN combines known interactions from multiple organisms, and has been used successfully to identify, for example, gene modules relevant in squamous carcinoma, based on a starting dataset of microarray data on differentially expressed genes between cancer cells and the surrounding tissue⁹⁸. Future studies applying such an Interolog PIN to the outcomes of genetic screens for developmental processes of interest could potentially overcome the problem of sparseness, as well as the biases towards proteins that are more heavily studied and thus better represented in organism-specific PINs.

3.4 Methods

3.4.1 Experimental model and subject details

Wild type and mutant lines of *Drosophila melanogaster* were obtained from publicly accessible stock centers and maintained as described in "Fly Stocks" below. Candidate genes were randomly assigned to batches for screening (see the Table C.1 for which genes were in each batch). F1 animals from the same cross were randomly assigned to experimental groups for phenotyping in all screens.

3.4.2 Fly stocks

Flies were reared at 25°C at 60% humidity with standard *Drosophila* food⁹⁹ containing yeast and in uncrowded conditions as previously defined²⁶. RNAi lines were obtained from the TRiP RNAi collection at the Bloomington *Drosophila* Stock Centre (BDSC) and from the Vienna *Drosophila* Resource Centre (VDRC). Oregon R was used as a wild type strain. The genotype of the *traffic jam:Gal4* line used in the screen was $y\ w; P\{w[+mW.hs] = GawB\}NP1624$ (Kyoto Stock Center, K104-055; abbreviated hereafter as *tj:Gal4*). The *hippo* RNAi line used in the screen was $y[1]\ v[1]; P\{y[+t7.7]v[+t1.8]=TRiP.HMS00006\}attP2$ (BDSC:33614; abbreviated hereafter as *hpo[RNAi]*).

3.4.3 Egg and ovariole number counts

Adult egg laying was quantified by crossing three virgin females of the desired genotype (see "Screen design" below) with two males in a vial containing standard food and yeast granules (day one) and then transferring them into a fresh food vial without yeast granules for a 24-hour period. Eggs from vials were then counted by visual inspection of the surface of the food in the vial. Males and females were transferred to fresh food vials without yeast granules, every day thereafter until day six. All egg laying measurements reported and analysed in the paper are the sum of the eggs laid by three adult female flies over the five days of this assay (days two through six without yeast granules). Data from any vial in which either a female or male died, during the course of the experiment, were not included in the analysis.

Ovariole number was quantified by mating ten virgin adult females with five virgin adult Oregon R males for three days post-eclosion in vials with yeast at 25°C and 60% humidity. After this three-day mating period, all 20 adult ovaries from the mated females were dissected in 1X PBS with 0.1% Triton-X-100 and stained with 1ug/ml Hoechst 33321 (1:10,000 of a 10mg/ml stock solution). Ovarioles were separated from each other with No. 5 forceps (Fine Science Tools) and counted by counting the number of germaria under a ZEISS Stemi 305 compact stereo microscope with a NIGHTSEA stereo microscope UV Fluorescence adaptor.

3.4.4 Screen design

In the primary screen (Figure 3.1a: *hpo*[RNAi] Egg Laying), 463 candidate genes (Table C.1) were screened for the effect of an RNAi-induced loss of gene function in a *hpo*[RNAi] background on the number of eggs laid in the first five days of mating (see "Egg and ovariole number counts" above) by adult females. These females were the F1 offspring of UAS:*candidate gene* RNAi males crossed to $P\{w[+mW.hs] = GawB\}NP1624; P\{y[+t7.7] v[+t1.8]=TRiP.HMS00006\}attP2 (tj:Gal4; UAS:hpo[RNAi])$ virgin adult females (Figure 3.1a: *hpo*[RNAi] Egg Laying). All genes that yielded an egg laying count with a $|Z_{gene}| > 1$ (see "Gene selection based on Z score and batch standardization" below) were selected to undergo two secondary screenings (n=273, Table 3.2; Figure 3.1d). First, these genes were screened for effects on the egg laying of mated adult female offspring from a cross of UAS:*candidate gene*[RNAi] males and *tj:Gal4* virgin adult females (Figure 3.1b: Egg Laying). Secondly, these genes were screened for effects on ovariole number in a *hpo*[RNAi] background. All 20 ovaries from ten adult female F1 offspring of a cross between UAS:*candidate gene*[RNAi] males to $P\{w[+mW.hs] = GawB\}NP1624; P\{y[+t7.7] v[+t1.8]=TRiP.HMS00006\}attP2 (tj:Gal4; UAS:hpo[RNAi])$ virgin adult females were scored for ovariole number (see "Egg and ovariole number counts" above). (Figure 3.1c: *hpo*[RNAi] Ovariole Number).

3.4.5 Gene selection based on Z score and batch standardization

Candidate genes were screened in batches with an average size of 50 genes. For each batch, control flies were the female F1 offspring of Oregon R males crossed to $P\{w[+mW.hs] = GawB\}NP1624; P\{y[+t7.7] v[+t1.8]=TRiP.HMS00006\}attP2 (tj:Gal4; UAS:hpo[RNAi])$ virgin adult

females. Because the control group in each batch had slightly different distributions of egg laying and ovariole number values (Figure C.1), it was inappropriate to compare absolute mean values between genes that were scored in different batches. Instead, comparisons of the Z score of each candidate (Z_{gene}) to its batch control group was used as a discriminant. This approach standardizes for batch effects and allows the comparison of all genotypes within and across the primary and secondary screens with a single metric (Z_{gene}).

Firstly, the mean and standard deviation of the eggs laid by the control genotype for a batch were calculated as μ_b and σ_b respectively. Then, using the number of eggs laid by adult females of a candidate gene RNAi (x_{gene}) of the same batch, the Z score for the egg laying count of that gene (Z_{gene}) was calculated as $Z_{gene} = (x_{gene} - \mu_b) / \sigma_b$. The same standardization protocol was applied to both egg laying and ovariole number counts of every gene and its corresponding batch control.

Ovariole numbers were derived from counts of the number of ovarioles per ovary for 20 ovaries per candidate gene, and a threshold of $|Z_{gene}| > 2$ was applied for ovariole number phenotype. Egg laying counts were derived from measurements of three females in a single vial per gene. We therefore chose to be more conservative in our Z score comparisons for the egg laying phenotype, than for ovariole number phenotype, and applied a stringent threshold of $|Z_{gene}| > 5$ to select genes of interest. All genes with $|Z_{gene}|$ values above these thresholds are referred to throughout the study as "positive candidates". (See Ipython notebooks 02_Z_score_calculation.ipynb and 02.2_Z_score_calculation_prediction.ipynb for code implementation and calculation of Z scores, and 06_Screen Analysis.ipynb for batch effects.)

3.4.6 Signalling pathway enrichment analysis

To study the enrichment of a particular signalling pathway in a group of candidate genes that had similar phenotypic effects revealed by the screen, custom scripts (see 07_Signaling_pathway_analysis.ipynb for code implementation) were generated to implement two different methods (Figure 3.3a, b; Figure C.2; Figure C.11a, b).

The first method is a numerical method that uses random sampling to calculate the null distribution of the number of members (M) of a signalling pathway (S) that would be expected at random in a set of genes of size (N). The script randomly sampled N genes from among the 463 tested *D. melanogaster* signalling genes 10,000 times, and counted the number of genes (M) that were members of the signalling pathway S. Positive candidates in each of the three screens were sorted by their presence in signalling pathways and counted. The Z score was then calculated by comparing the experimentally observed number of positive candidates in each signalling pathway against the randomly sampled null distribution.

The second method used the hypergeometric p-value to calculate the probability of M members of a signalling pathway being in a group of N genes, given a starting population of 463 tested *D. melanogaster* signalling genes, and the known attribution to a pathway S of each gene.

3.4.7 Protein-Protein Interaction Network (PIN) building

There is no standard complete Protein-Protein Interaction network (PIN) available for *Drosophila melanogaster*. However, there exist many smaller networks from different screens, as well as literature extractions. We therefore combined data from these sources and then created a PIN for use in the present study, as follows:

Step 1: Several screens assessing protein-protein interactions have been centralized in a database called DroID: <http://www.droidb.org>. The version DroID_v2018_08 was used. All available datasets were first downloaded from that database using this link:

<http://www.droidb.org/Downloads.jsp>. The description of all of these datasets can be found here: <http://www.droidb.org/DBdescription.jsp>

Step 2: We used the datasets from all screens that assessed direct protein-protein interactions and did not use the interolog database (predicted protein interaction based on mouse human and yeast PPIs). These direct assessment screens were seven in total, as follows:

- [Finley Yeast Two-Hybrid Data](#) (size 2.0 MB; 3610 Nodes & 9007 Edges)
- [Curagen Yeast Two-Hybrid Data](#) (size 4.6 MB; 6678 Nodes & 19506 Edges)

- [Hybrigenics Yeast Two-Hybrid Data](#) (size 381 KB; 1269 Nodes & 1842 Edges)
- [Perrimon co-AP complex](#) (size 108 KB; 252 Nodes & 384 Edges)
- [DPiM co-AP complex](#) (size 6.3 MB; 3732 Nodes & 17652 Edges)
- [PPI from other databases](#) (size 16.2 MB; 7524 Nodes & 47471 Edges)
- [PPI curated by FlyBase](#) (size 7.4 MB; 5125 Nodes & 31491 Edges)

We did not consider self-loop edges from proteins predicted to interact with themselves (homotypic or self-interactions). An important element to note is that the PPIs curated by FlyBase is a literature-based PPIs. FlyBase protein-protein interactions are experimentally derived physical interactions curated from the literature by FlyBase and does not include FlyBase-curated genetic interactions.

Step 3: We concatenated the seven datasets listed above into a single unique database. A custom python script was created that downloads and reads each of the above seven unique PPI tables, and generates a single PIN (see 01_PIN_builder.ipynb). From this concatenation, a single edge undirected network was created and saved. This network is hereafter referred to as the PIN (see 01_PIN_builder.ipynb). The PIN contains 10,632 proteins (nodes) and 85,019 interactions (edges), giving a network density of 0.0015.

3.4.8 Network metric computations

The centrality of a node is often used as a measure of a node's importance in a network. Within a PIN, the centrality of a gene reflects the number of interactions in which the gene directly or indirectly participates. Four different centrality metrics were computed for all genes in the PIN using the networkx python library:

- (1) **Betweenness** reflects the number of shortest paths passing through a gene.
- (2) **Eigenvector** is a measure of the influence of a gene in the network.
- (3) **Closeness** measures the sum of shortest distance of a gene to all the other genes.
- (4) **Degree centrality** corresponds to the normalized number of edges of a gene in the network.

While there exist more centrality measures, these four are commonly used to assess biological networks. These computed centrality parameters of the genes measured in the screen were computed with `03_ROC_curve_analysis_of_network_metrics.ipynb`, and are reported in the Table C.1 (see `09_Making_the_database_table.ipynb`).

3.4.9 Receiver Operating Characteristic (ROC) curves

To check whether the centrality of a gene in the network could predict the phenotypic effect produced by RNAi against that gene, ROC curves were plotted for the four aforementioned centrality measures of each gene in each screen. A ROC analysis is used to measure the correlation between a continuous variable (centrality) and a binary outcome (above or below Z score threshold). Therefore, for each screen, measured genes were rank ordered from high centrality to low centrality, and plotted against the binary outcome of $|Z_{gene}|$ being above or below the appropriate $|Z \text{ score}|$ threshold (>5 for egg laying and >2 for ovariole number). The Area Under the Curve (AUC) measures the extent of correlation between centrality and effect of a gene on measured phenotype. AUC above or below 0.5 indicates a positive or negative correlation respectively, while an AUC of 0.5 indicates no correlation of the parameters. The scikit-learn python package was used to calculate the AUC of each ROC curve plotted (see `03_ROC_curve_analysis_of_network_metrics.ipynb`).

3.4.10 Building degree-controlled randomized networks

We assessed the modularity of the networks by comparing the network metrics of each sub-network to a degree-controlled randomly sampled network. To generate this degree controlled random network, we applied a previously developed method¹⁰⁰. In short, nodes in the PPI are binned by degree with the minimum size of each bin being set at 100 nodes. Bins are constructed iteratively from the lowest degree to the highest degree in the network. To sample a set of nodes, the sub-network degree distribution is computed, using the bin cut-off, from the PPI. Then, nodes are randomly selected from each bin to match this degree distribution (see `05.2_Degree_Controlled_Testing.ipynb` for code implementation).

3.4.11 Assessing the utility of the Seed Connector Algorithm in building network modules

Network modules were assessed using the previously published Seed Connector Algorithm (SCA)^{66,67}, implemented here in python (see 04_Seed-Connector.ipynb) and illustrated in Figure 3.5a. Creating a module using the SCA requires a list of seed genes and a PIN. From each of the three screens, we selected the genes whose $|Z_{gene}|$ value was above the threshold and created three seed lists respectively (Figure 3.4c: Egg laying, *hpo[RNAi]* egg laying and *hpo[RNAi]* ovariole ‘seed’ list). A fourth list consisting of the intersection of the aforementioned seed lists was also collated and called the core ‘seed’ list (Figure 3.4b). Genes were assigned in the core list if they passed the Z threshold in all 3 screens. The Seed Connector Algorithm was then executed on each of these seed lists using the PIN. Not all genes in the four seed lists were found in the PIN (specifically, CG12147 in the *hpo[RNAi]* Egg Laying seed list and CG6104 in the *hpo[RNAi]* Ovariole number seed list were absent from the PIN) and were therefore eliminated from further network analysis. The removal of these two genes accounts for the variation in the number of positive candidates in Table 3.2 and the number of seed genes in the module. Modules were obtained for each seed list (Figure 3.5b; Figures C.6 to C.8) consisting of the seed genes (circles in Figure 3.5b and Figures C.6 to C.8) and previously untested genes added by the SCA (squares in Figure 3.5b and Figures C.6 to C.8) to increase the LCC size that we refer to as connector genes (see 04_Seed-Connector.ipynb). The results of the algorithm are summarized in the Table C.1.

The modularity of the sub-networks was then assessed using four network metrics namely Largest Connected Component (LCC), number of edges, network density and average shortest path in the LCC. Each metric for each module was assessed using distance of the network metric to a null distribution. Initially, the null distribution was calculated by taking 1000 samples of 463 genes randomly selected from the PIN and calculating the above metrics. We found that the 463 genes selected in the signalling screen were already more connected than the null distribution of sets of 463 genes randomly selected from the PIN (Figure C.9a). Therefore, to avoid a false positive detection of modularity, the four experimentally obtained sub-networks were compared

to null distributions obtained by randomly sampling an equal number of genes from the 463 signalling candidate genes selected for our screen. For each of the four modules, comparison of the metrics was performed on the seed lists and the sub-network after the SCA. Most metrics were enriched in the seed group when compared to the null distribution with the exception of the Average shortest path (Figure C.9b, light red line). The sub-networks obtained from the SCA further increased all four metrics suggesting the modularity of the four sub-networks (Figure C.9b, dark red line; see 05_Network_Module_testing.ipynb for code implementation).

3.4.12 Meta network

To build the meta network, the genes from all four sub-networks were concatenated into one network. This network was then visually sorted in an approach akin to projecting the network onto a Venn Diagram. The meta network was sorted by which of the three screens the gene was positive in. The intersections were genes whose $|Z_{gene}|$ value was above the threshold in more than one and possibly all three of the screening paradigms. For example, if a gene was found in the *hpo[RNAi]* Ovariole Number and Egg Laying sub-networks it is then assigned to the dual positive group *hpo[RNAi]* Ovariole Number / Egg Laying (Figure 3.6a, sub-network VI). After applying this grouping strategy, the connectivity across the groups was studied by calculating the edge density between all groups ($density = \frac{Edges_{s_1,2}}{Nodes_1 * Nodes_2}$). Finally, the proportion of each signalling candidate in each of those groups was calculated by taking the number of members of a signalling pathway divided by the total members of a group (see Ipython notebook 08_MetaModule_Analysis.ipynb). A single gene, *sloppy paired 1*, was a seed in the Egg Laying sub-network and also a connector in the *hpo[RNAi]* Egg Laying sub-network; it fell within sub-network VII in the meta network, and is marked as a seed (grey) in Figure 3.6a.

3.4.13 Number of samples

The number of samples across the different screens were as follows:

***hpo[RNAi]* Egg Laying and Egg Laying screens**

- Controls: five vials of three females and two males
- Sample: one vial of three females and two males

***hpo[RNAi]* Ovariole number screen**

- Controls: 20 flies, two ovaries per fly considered as independent measurements
- Sample: 10 flies, two ovaries per fly considered as independent measurements

3.4.14 Correction of batch effect

Despite best efforts to maintain the exact same condition between each experiment, some variation was measured between the batches. Control flies showed variations in both measured phenotypes, ovariole number and egg laying (Figure C.1). In order to compare the values measured across different batches, each sample was standardized by calculating its Z score (Z_{gene}) to the control distribution. For each batch, the measurements for controls were pooled into a distribution, and the mean and standard deviation was computed. Then each sample was compared to its respective batch and its Z score computed (see "*Gene selection based on Z score and batch standardization*" for formula).

3.4.15 Statistical analysis

All statistical analyses were performed using the scipy stats module (<https://www.scipy.org/>) and scikit-learn (<https://scikit-learn.org/>) python package. Statistical tests and p-values are reported in the figure legends. All statistical tests can be found in the Ipython notebooks mentioned below.

3.4.16 Data and code availability

This study generated a series of python3 Ipython notebook files that perform the entire analysis presented in this study. All the results presented in this paper, including the figures with the exception of the network visualizations, which were created using Cytoscape3 (<https://cytoscape.org/>) can be reproduced by running the aforementioned python3 code. The raw data, calculations made with these data, and code used for calculations and analyses (Ipython notebooks) are available as supplementary information. For ease of access, legibility and reproducibility, the code and datasets have been deposited in a GitHub repository available at

[https://github.com/extavourlab/hpo_ova_eggL_screen.](https://github.com/extavourlab/hpo_ova_eggL_screen)

References

- [1] X. Yang and T. Xu. Molecular mechanism of size control in development and human diseases. *Cell Res.*, 21(5):715–729, May 2011.
- [2] M. P. Kambysellis and W. B. Heed. Studies of Oogenesis in Natural Populations of *Drosophilidae*. I. Relation of Ovarian Development and Ecological Habitats of the Hawaiian Species. *Am. Nat.*, 105(941):31–49, January 1971.
- [3] R. C. King. *Ovarian Development in Drosophila melanogaster*. Academic Press, New York, 1970.
- [4] T. A. Markow, S. Beall, and L. M. Matzkin. Egg size, embryonic development time and ovoviviparity in *Drosophila* species. *J. Evol. Biol.*, 22(2):430–434, February 2009.
- [5] D. P. Sarikaya, S. H. Church, L. P. Lagomarsino, K. N. Magnacca, S. L. Montgomery, D. K. Price, K. Y. Kaneshiro, and C. G. Extavour. Reproductive Capacity Evolves in Response to Ecology through Common Changes in Cell Number in Hawaiian *Drosophila*. *Curr. Biol.*, 29(11):1877–1884.e6, June 2019.
- [6] P. Klepsatel, M. Gálíková, N. De Maio, S. Ricci, C. Schlötterer, and T. Flatt. Reproductive and post-reproductive life history of wild-caught *Drosophila melanogaster* under laboratory conditions. *J. Evol. Biol.*, 26(7):1508–1520, July 2013.
- [7] S. R'kha, B. Moreteau, J. A. Coyne, and J. R. David. Evolution of a lesser fitness trait: egg production in the specialist *Drosophila sechellia*. *Genetical Research*, 69(1):17–23, 1997.

- [8] R. B. R. Azevedo, V. French, and L. Partridge. Thermal evolution of egg size in *Drosophila melanogaster*. *Evolution*, 50(6):2338–2345, December 1996.
- [9] P. Capy, E. Pla, and J. R. David. Phenotypic and genetic variability of morphometrical traits in natural populations of *Drosophila melanogaster* and *D. simulans*. II. Within-population variability. *Genet. Sel. Evol.*, 26(1):15, February 1994.
- [10] J. R. David and C. Bocquet. Similarities and differences in latitudinal adaptation of two *Drosophila* sibling species. *Nature*, 257(5527):588–590, October 1975.
- [11] P. Capy, E. Pla, and J. R. David. Phenotypic and genetic variability of morphometrical traits in natural populations of *Drosophila melanogaster* and *D. simulans*. I. Geographic variations. *Genet. Sel. Evol.*, 25(6):517, December 1993.
- [12] P. Klepsatel, M. Gálíková, N. De Maio, C. D. Huber, C. Schlötterer, and T. Flatt. Variation in thermal performance and reaction norms among populations of *Drosophila melanogaster*. *Evolution*, 67(12):3573–3587, December 2013.
- [13] S. R’Kha, P. Capy, and J. R. David. Host-plant specialization in the *Drosophila melanogaster* species complex: a physiological, behavioral, and genetical analysis. *Proc. Natl. Acad. Sci. U. S. A.*, 88(5):1835–1839, March 1991.
- [14] A. O. Bergland, A. Genissel, S. V. Nuzhdin, and M. Tatar. Quantitative trait loci affecting phenotypic plasticity and the allometric relationship of ovariole number and thorax length in *Drosophila melanogaster*. *Genetics*, 180(1):567–582, September 2008.
- [15] A. S. Lobell, R. R. Kaspari, Y. L. Serrano Negrón, and S. T. Harbison. The Genetic Architecture of Ovariole Number in *Drosophila melanogaster*: Genes with Major, Quantitative, and Pleiotropic Effects. *G3*, 7(7):2391–2403, July 2017.
- [16] V. Orgogozo, K. W. Broman, and D. L. Stern. High-resolution quantitative trait locus mapping reveals sign epistasis controlling ovariole number between two *Drosophila* species. *Genetics*, 173(1):197–205, May 2006.
- [17] M. L. Wayne and L. M. McIntyre. Combining mapping and arraying: An approach to candidate gene identification. *Proc. Natl. Acad. Sci. U. S. A.*, 99(23):14903–14906, November

- 2002.
- [18] M. L. Wayne, J. B. Hackett, and T. F. C. Mackay. QUANTITATIVE GENETICS OF OVARIOLE NUMBER IN *DROSOPHILA MELANOGASTER*. I. SEGREGATING VARIATION AND FITNESS. *Evolution*, 51(4):1156–1163, August 1997.
- [19] M. L. Wayne, J. B. Hackett, C. L. Dilda, S. V. Nuzhdin, E. G. Pasyukova, and T. F. Mackay. Quantitative trait locus mapping of fitness-related traits in *Drosophila melanogaster*. *Genet. Res.*, 77(1):107–116, February 2001.
- [20] R. C. King, S. K. Aggarwal, and U. Aggarwal. The development of the female *Drosophila* reproductive system. *J. Morphol.*, 124(2):143–166, February 1968.
- [21] D. Godt and F. A. Laski. Mechanisms of cell rearrangement and cell recruitment in *Drosophila* ovary morphogenesis and the requirement of bric à brac. *Development*, 121(1):173–187, January 1995.
- [22] R. Keller. Mechanisms of elongation in embryogenesis. *Development*, 133(12):2291–2302, June 2006.
- [23] I. Sahut-Barnola, B. Dastugue, and J.-L. Couderc. Terminal filament cell organization in the larval ovary of *Drosophila melanogaster*: ultrastructural observations and pattern of divisions. *Roux's Arch. Dev. Biol.*, 205(7-8):356–363, May 1996.
- [24] I. Sahut-Barnola, D. Godt, F. A. Laski, and J. L. Couderc. *Drosophila* ovary morphogenesis: analysis of terminal filament formation and identification of a gene required for this process. *Dev. Biol.*, 170(1):127–135, July 1995.
- [25] J. Chen, D. Godt, K. Gunsalus, I. Kiss, M. Goldberg, and F. A. Laski. Cofilin/ADF is required for cell motility during *Drosophila* ovary development and oogenesis. *Nat. Cell Biol.*, 3(2):204–209, February 2001.
- [26] D. P. Sarikaya and C. G. Extavour. The Hippo pathway regulates homeostatic growth of stem cell niche precursors in the *Drosophila* ovary. *PLoS Genet.*, 11(2):e1004962, February 2015.

- [27] D. Hilman and U. Gat. The evolutionary history of YAP and the *hippo*/YAP pathway. *Mol. Biol. Evol.*, 28(8):2403–2417, August 2011.
- [28] A. Sebé-Pedrós, Y. Zheng, I. Ruiz-Trillo, and D. Pan. Premetazoan origin of the *hippo* signaling pathway. *Cell Rep.*, 1(1):13–20, January 2012.
- [29] E. T. Ables and D. Drummond-Barbosa. Steroid Hormones and the Physiological Regulation of Tissue-Resident Stem Cells: Lessons from the *Drosophila* Ovary. *Curr Stem Cell Rep*, 3(1):9–18, March 2017.
- [30] L. Gilboa. Organizing stem cell units in the *Drosophila* ovary. *Curr. Opin. Genet. Dev.*, 32: 31–36, June 2015.
- [31] D. A. Green, 2nd, D. P. Sarikaya, and C. G. Extavour. Counting in oogenesis. *Cell Tissue Res.*, 344(2):207–212, May 2011.
- [32] L. LaFever and D. Drummond-Barbosa. Direct control of germline stem cell division and cyst growth by neural insulin in *Drosophila*. *Science*, 309(5737):1071–1073, August 2005.
- [33] L. LaFever, A. Feoktistov, H.-J. Hsu, and D. Drummond-Barbosa. Specific roles of Target of rapamycin in the control of stem cells and their progeny in the *Drosophila* ovary. *Development*, 137(13):2117–2126, July 2010.
- [34] D. Gancz and L. Gilboa. Insulin and Target of rapamycin signaling orchestrate the development of ovarian niche-stem cell units in *Drosophila*. *Development*, 140(20): 4145–4154, October 2013.
- [35] D. A. Green, 2nd and C. G. Extavour. Convergent evolution of a reproductive trait through distinct developmental mechanisms in *Drosophila*. *Dev. Biol.*, 372(1):120–130, December 2012.
- [36] H.-J. Hsu and D. Drummond-Barbosa. Insulin levels control female germline stem cell maintenance via the niche in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.*, 106(4):1117–1121, January 2009.

- [37] D. Gancz, T. Lengil, and L. Gilboa. Coordinated regulation of niche and stem cell precursors by hormonal signaling. *PLoS Biol.*, 9(11):e1001202, November 2011.
- [38] J. Hodin and L. M. Riddiford. The ecdysone receptor and ultraspiracle regulate the timing and progression of ovarian morphogenesis during *Drosophila* metamorphosis. *Dev. Genes Evol.*, 208(6):304–317, August 1998.
- [39] J. Hodin and L. M. Riddiford. Parallel alterations in the timing of ovarian ecdysone receptor and ultraspiracle expression characterize the independent evolution of larval reproduction in two species of gall midges (Diptera: Cecidomyiidae). *Dev. Genes Evol.*, 210(7):358–372, July 2000.
- [40] M. W. Gonzalez and M. G. Kann. Chapter 4: Protein interactions and disease. *PLoS Comput. Biol.*, 8(12):e1002819, December 2012.
- [41] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.
- [42] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.
- [43] R. Albert and A.-L. Barabási. *Statistical Mechanics of Complex Networks*, volume 74. Springer, Berlin, Heidelberg, 2003.
- [44] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–113, February 2004.
- [45] S. I. Berger, J. M. Posner, and A. Ma'ayan. Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics*, 8:372, October 2007.
- [46] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrola, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong,

- C. A. Stanyon, R. L. Finley, Jr, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736, December 2003.
- [47] L. S. Gramates, S. J. Marygold, G. D. Santos, J. M. Urbano, G. Antonazzo, B. B. Matthews, A. J. Rey, C. J. Tabone, M. A. Crosby, D. B. Emmert, K. Falls, J. L. Goodman, Y. Hu, L. Ponting, A. J. Schroeder, V. B. Strelets, J. Thurmond, P. Zhou, and C. FlyBase. FlyBase at 25: looking to the future. *Nucleic Acids Res.*, 45:663– 671, 2016.
- [48] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, 38(Database issue):D355–60, January 2010.
- [49] A. Mbodj, G. Junion, C. Brun, E. E. M. Furlong, and D. Thieffry. Logical modelling of *Drosophila* signalling pathways. *Mol. Biosyst.*, 9(9):2248–2258, September 2013.
- [50] T. Ideker and R. Sharan. Protein networks in disease. *Genome Res.*, 18(4):644–652, April 2008.
- [51] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.
- [52] M. W. Hahn and A. D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.*, 22(4):803–806, April 2005.
- [53] A. Ma’ayan. Introduction to network analysis in systems biology. *Sci. Signal.*, 4(190):tr5, September 2011.
- [54] M. Jalili, A. Salehzadeh-Yazdi, S. Gupta, O. Wolkenhauer, M. Yaghmaie, O. Resendis-Antonio, and K. Alimoghaddam. Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. *Front. Physiol.*, 7:375, August 2016.
- [55] D. Koschützki and F. Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul. Syst. Bio.*, 2:193–201, May 2008.

- [56] L. J. Foster, C. L. de Hoog, Y. Zhang, Y. Zhang, X. Xie, V. K. Mootha, and M. Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125(1):187–199, April 2006.
- [57] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9(12):3273–3297, December 1998.
- [58] B. S. Srinivasan, N. H. Shah, J. A. Flannick, E. Abeliuk, A. F. Novak, and S. Batzoglou. Current progress in network research: toward reference networks for key model organisms. *Brief. Bioinform.*, 8(5):318–332, September 2007.
- [59] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52, December 1999.
- [60] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, August 2002.
- [61] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, April 2004.
- [62] K. D. Bromberg, A. Ma’ayan, S. R. Neves, and R. Iyengar. Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. *Science*, 320(5878):903–909, May 2008.
- [63] S.-S. C. Huang and E. Fraenkel. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci. Signal.*, 2(81):ra40, July 2009.
- [64] I. Albert and R. Albert. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346–3352, December 2004.
- [65] H. Yu, A. Paccanaro, V. Trifonov, and M. Gerstein. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7):823–829, April 2006.

- [66] R.-S. Wang, K. T. Hall, F. Giulianini, D. Passow, T. J. Kaptchuk, and J. Loscalzo. Network analysis of the genomic basis of the placebo effect. *JCI Insight*, 2(11):93911, June 2017.
- [67] R.-S. Wang and J. Loscalzo. Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications. *J. Mol. Biol.*, 430(18 Pt A): 2939–2950, September 2018.
- [68] J. Y. Chen, C. Shen, and A. Y. Sivachenko. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac. Symp. Biocomput.*, 11:367–378, 2006.
- [69] G. Gonzalez, J. C. Uribe, L. Tari, C. Brophy, and C. Baral. Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. *Pac. Symp. Biocomput.*, 12:28–39, 2007.
- [70] Y. Cohet and J. David. Control of the adult reproductive potential by preimaginal thermal conditions : A study in *Drosophila melanogaster*. *Oecologia*, 36(3):295–306, January 1978.
- [71] J. Hodin and L. M. Riddiford. Different mechanisms underlie phenotypic plasticity and interspecific variation for a reproductive character in drosophilids (Insecta: Diptera). *Evolution*, 54(5):1638–1653, October 2000.
- [72] S. Llamazares, A. Moreira, A. Tavares, C. Girdham, B. A. Spruce, C. Gonzalez, R. E. Karess, D. M. Glover, and C. E. Sunkel. polo encodes a protein kinase homolog required for mitosis in *Drosophila*. *Genes Dev.*, 5(12A):2153–2165, December 1991.
- [73] M. C. Edwards, C. Wong, and S. J. Elledge. Human cyclin K, a novel RNA polymerase II-associated cyclin possessing both carboxy-terminal domain kinase and Cdk-activating kinase activity. *Mol. Cell. Biol.*, 18(7):4291–4300, July 1998.
- [74] T. Ito, M. Bulger, R. Kobayashi, and J. T. Kadonaga. *Drosophila* NAP-1 is a core histone chaperone that functions in ATP-facilitated assembly of regularly spaced nucleosomal arrays. *Mol. Cell. Biol.*, 16(6):3112–3124, June 1996.
- [75] A. T. Chao, H. A. Dierick, T. M. Addy, and A. Bejsovec. Mutations in eukaryotic release factors 1 and 3 act as general nonsense suppressors in *Drosophila*. *Genetics*, 165(2):601–612, October 2003.

- [76] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, October 2008.
- [77] R. T. H. Lee, Z. Zhao, and P. W. Ingham. Hedgehog signalling. *Development*, 143(3):367–372, February 2016.
- [78] C.-M. Lai, K.-Y. Lin, S.-H. Kao, Y.-N. Chen, F. Huang, and H.-J. Hsu. Hedgehog signaling establishes precursors for germline stem cell niches by regulating cell adhesion. *J. Cell Biol.*, 216(5):1439–1453, May 2017.
- [79] A. E. Christiansen, T. Ding, and A. Bergmann. Ligand-independent activation of the Hedgehog pathway displays non-cell autonomous proliferation during eye development in *Drosophila*. *Mech. Dev.*, 129(5-8):98–108, July 2012.
- [80] K. Jiang, Y. Liu, J. Zhang, and J. Jia. An intracellular activation of Smoothed that is independent of Hedgehog stimulation in *Drosophila*. *J. Cell Sci.*, 131(1):211367, January 2018.
- [81] J. Bolívar, J. Pearson, L. López-Onieva, and A. González-Reyes. Genetic dissection of a stem cell niche: the case of the *Drosophila* ovary. *Dev. Dyn.*, 235(11):2969–2979, November 2006.
- [82] J.-L. Couderc, D. Godt, S. Zollman, J. Chen, M. Li, S. Tiong, S. E. Cramton, I. Sahut-Barnola, and F. A. Laski. The bric à brac locus consists of two paralogous genes encoding BTB/POZ domain proteins and acts as a homeotic and morphogenetic regulator of imaginal development in *Drosophila*. *Development*, 129(10):2419–2433, May 2002.
- [83] D. Godt, J. L. Couderc, S. E. Cramton, and F. A. Laski. Pattern formation in the limbs of *Drosophila*: bric à brac is expressed in both a gradient and a wave-like pattern and is

- required for specification and proper segmentation of the tarsus. *Development*, 119(3):799–812, November 1993.
- [84] A. W. Allbee, D. E. Rincon-Limas, and B. Biteau. Lmx1a is required for the development of the ovarian stem cell niche in *Drosophila*. *Development*, 145(8):163394, April 2018.
- [85] L. Dobens, A. Jaeger, J. S. Peterson, and L. A. Raftery. Bunched sets a boundary for Notch signaling to pattern anterior eggshell structures during *Drosophila* oogenesis. *Dev. Biol.*, 287(2):425–437, November 2005.
- [86] P. Cayirlioglu, W. O. Ward, S. C. Silver Key, and R. J. Duronio. Transcriptional repressor functions of *Drosophila* E2F1 and E2F2 cooperate to inhibit genomic DNA synthesis in ovarian follicle cells. *Mol. Cell. Biol.*, 23(6):2123–2134, March 2003.
- [87] K. Barrett, M. Leptin, and J. Settleman. The Rho GTPase and a putative RhoGEF mediate a signaling pathway for the cell shape changes in *Drosophila* gastrulation. *Cell*, 91(7):905–915, December 1997.
- [88] T. Mizuno, M. Amano, K. Kaibuchi, and Y. Nishida. Identification and characterization of *Drosophila* homolog of Rho-kinase. *Gene*, 238(2):437–444, October 1999.
- [89] K. E. Kasza, D. L. Farrell, and J. A. Zallen. Spatiotemporal control of epithelial remodeling by regulated myosin phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.*, 111(32):11732–11737, August 2014.
- [90] A. J. Ridley. Rho GTPases and actin dynamics in membrane protrusions and vesicle trafficking. *Trends Cell Biol.*, 16(10):522–529, October 2006.
- [91] L. Lewellyn, M. Cetera, and S. Horne-Badovinac. Misshapen decreases integrin levels to promote epithelial motility and planar polarity in *Drosophila*. *J. Cell Biol.*, 200(6):721–729, March 2013.
- [92] T. J. C. Harris and M. Peifer. Adherens junction-dependent and -independent steps in the establishment of epithelial cell polarity in *Drosophila*. *J. Cell Biol.*, 167(1):135–147, October 2004.

- [93] R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, 12(1):37–46, January 2002.
- [94] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.
- [95] X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics*, 11 Suppl 1:S3, February 2010.
- [96] Y. Zhang, H. Lin, Z. Yang, J. Wang, and Y. Liu. An uncertain model-based approach for identifying dynamic protein complexes in uncertain protein-protein interaction networks. *BMC Genomics*, 18(Suppl 7):743, October 2017.
- [97] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or “interologs”. *Genome Res.*, 11(12):2120–2126, December 2001.
- [98] S. Wachi, K. Yoneda, and R. Wu. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 21(23):4205–4208, December 2005.
- [99] D. P. Sarikaya, A. A. Belay, A. Ahuja, A. Dorta, D. A. Green, 2nd, and C. G. Extavour. The roles of cell size and cell number in determining ovariole number in *Drosophila*. *Dev. Biol.*, 363(1):279–289, March 2012.
- [100] E. Guney, J. Menche, M. Vidal, and A.-L. Barábasi. Network-based in silico drug efficacy screening. *Nat. Commun.*, 7:10331, February 2016.

The student of heredity is confronted with two different groups of problems. The first group, including such questions as: 'Why do two white mice produce another white mouse, but not a black or a brown one?' has been answered, at least up to a point, by the geneticists. The second, typified by such a question as: 'Why do two mice produce a mouse, and not a rabbit, a mass of Protozoa, or a sarcoma?' has been answered much less satisfactorily.

J. B. S. Haldane, 1940

4

Studying germ layer specification in the early embryo of *Parhyale hawaiiensis* with single-cell RNA sequencing

ABSTRACT

Understanding the molecular mechanisms underlying the specification of the germ layers is a fundamental question of developmental biology. In this chapter, I focused on measuring and understanding the gene expression profiles of the cells generated during the first three days of *P. hawaiiensis* embryogenesis. To achieve a comprehensive understanding of the differentiation processes underlying germ layers specification, I aimed to generate developmental time course single-cell RNA sequencing libraries. I developed cell separation protocols required to sequence the transcriptomes of single cells with Cel-Seq2 and inDrop. After analyzing the initial libraries generated, it appeared that the sequencing depth of the cells was not sufficient to study the specification of germ layers. However, I demonstrate the technical feasibility of generating single-cell RNA sequencing data for developing *P. hawaiiensis* embryos.

CONTRIBUTIONS

For this study, I received the help of members of Allon Klein's laboratory, in particular James Briggs who shared his dissociation protocol. Moreover, droplet capture and sequencing were performed by the members of the Harvard Medical School Single Cell Core and the Harvard Bauer Core Facility. Cassandra G. Extavour assisted and taught me how to separate the blastomeres of *P. hawaiiensis* embryos.

ACKNOWLEDGEMENT

I would like to thank Allon Klein and the members of the Harvard Medical School Single Cell Core for the help and advice they offered throughout my Ph.D. I would also like to thank Claire Reardon of the Harvard Bauer Core Facility for her help in sequencing the libraries and advice with the CelSeq2 protocol. Finally I would like to thank Shreeharsha (Harsha) Tarikere from the Extavour laboratory for the long discussions and exchanges regarding the dissociation of complex tissues in single cells and their subsequent sequencing.

4.1 Introduction

In the next two chapters, I present the work that I have conceptualized and performed towards understanding the early specification of germ layers in the crustacean amphipod *P. hawaiiensis*. In this chapter, the central question I tried to tackle was: What changes in transcriptomic profiles does *P. hawaiiensis* undergo during the specification and differentiation of its germ layers? I introduce the emerging model organism I used and explain the choice of this organism. I continue by presenting what is known about the specification of germ layers in this organism, and place this within a broader context. I introduce the original hypothesis and plans that guided the experimental design. I conclude by presenting the first part of this project containing the presentation of the single-cell RNA sequencing techniques I employed and the results obtained.

4.1.1 The amphipod *P. hawaiiensis*, an emerging model organism for the study of germ layer specification

P. hawaiiensis is a small crustacean amphipod measuring between 5 and 15 mm as an adult¹. It has a circumtropical, worldwide, intertidal and shallow water ecology, and follows a detritivorous diet². It is commonly found in mangroves where it participates in the destruction of leafy materials³, and recent studies of its genome showed the presence of multiple lignocellulose digestive enzymes⁴. *P. hawaiiensis* has been found to thrive in areas of rapid salinity and temperature changes^{2,3}. It can also live at high population density, with upwards of thousands of individuals in one square meter^{2,3}. The female reproductive cycle is two weeks long, where each cycle leads to a molt and a new egg brood (Figure 4.1). All those factors make *P. hawaiiensis* a very resilient animal with a widespread ecology and fast generation time, traits that are particularly favorable to the establishment of laboratory culture. The laboratory of Nipam Patel established the first colony from individuals coming from the filtration system of the Chicago aquarium⁵. The first embryonic divisions are holoblastic, allowing the injection of mRNA or other molecules into each blastomere⁶. The early divisions are highly stereotypical, leading to a reproducible position of each blastomere by the 8 and 16 cell stages⁷ (Figure 4.1).

The fate of each cell is established as early as the 8 cell stage, where each of the 8 blastomeres will give rise to unique cell lineages (Figure 4.1). Three blastomeres will give rise to the Ectoderm, three to the Mesoderm, one to the Endoderm, and one to the Germline^{1,7} (Figure 4.1). Similar to other arthropod species, *P. hawaiiensis* embryos develop a germ band that elongates along the anterior-posterior axis and will become divided into the different segments of the adult organism^{6,8}. Moreover, the ectodermal cells are organized in a grid-like pattern at the germ band stage, where each row of the grid follows a stereotypical division cycle^{6,9}. Importantly, the genome and multiple transcriptomes of this organism have been sequenced^{4,10,11}, and the asymmetrical inheritance of maternal transcripts was measured at the S4 stage (8 cells)¹⁰, which could be necessary for the early specification of cell fate¹⁰. Finally, the relatively low number of cells during the early stages of embryogenesis (1200-1500 cells by the 3rd day of development) make this organism well suited for the tracking of cell behavior^{7,9}.

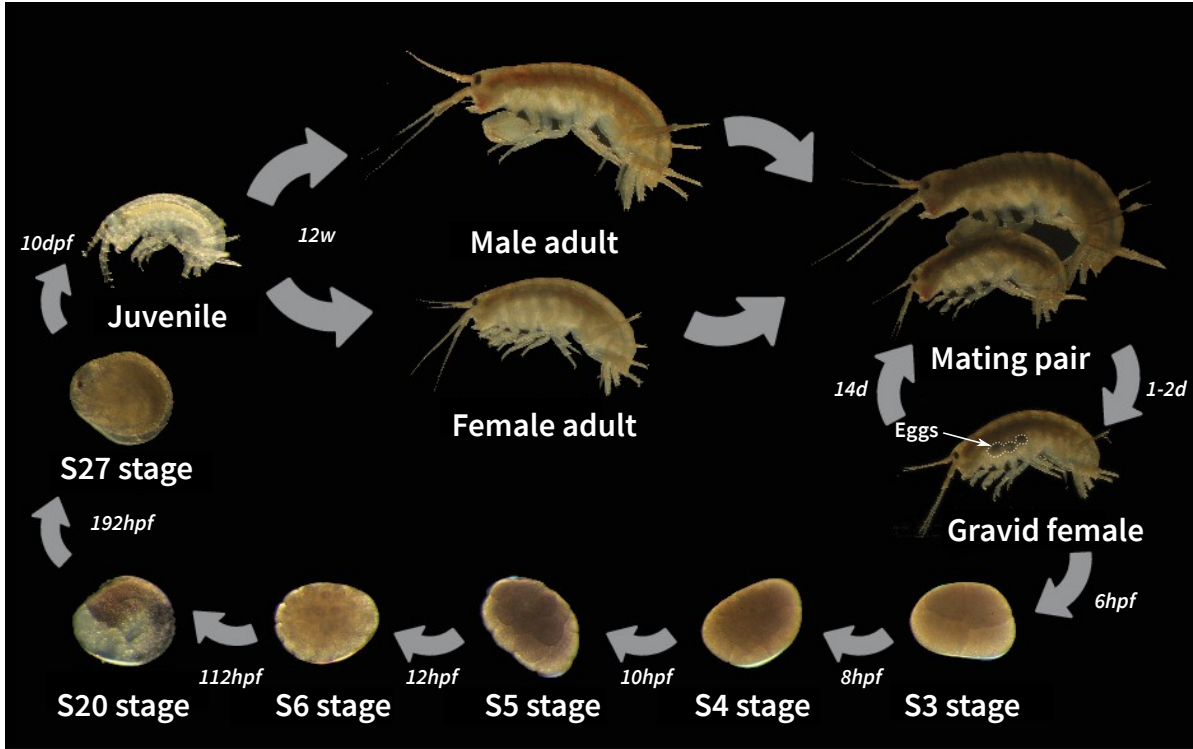
In addition to the multiple advantages this organism offers to understand the development of amphipods and crustaceans, it also shows high regenerative capacity (reviewed in Sun and Patel¹²). After ablation, an adult limb can be regenerated¹³ through the induction and migration of progenitor cells¹⁴. Moreover, at the early stages of development, the ablation of an ectodermal or mesodermal blastomere can be compensated for through an intra-germ layer compensation mechanism¹⁵ (Figures 4.2 and 4.3). This phenomenon might be considered counter-intuitive for this organism given the high degree of stereotypic developmental processes^{1,6,7}. To conclude, thanks to the adaptation of multiple biological techniques, *P. hawaiiensis* has become an attractive crustacean arthropod for the study of development.

During the early phase of the development of metazoans, cells differentiate into three populations, called germ layers. The Ectoderm gives rise to the external's most tissues, the Endoderm to the internal, often luminal, tissues, and the Mesoderm gives rise to tissues generally situated between the other two (reviewed in Gerberding and Patel¹⁶). However, while the gene regulatory network underlying germ layer specification has been studied in a number of traditional model organisms^{17,18,19,20}, these processes are less understood in other animals (discussed in Chapter 0).

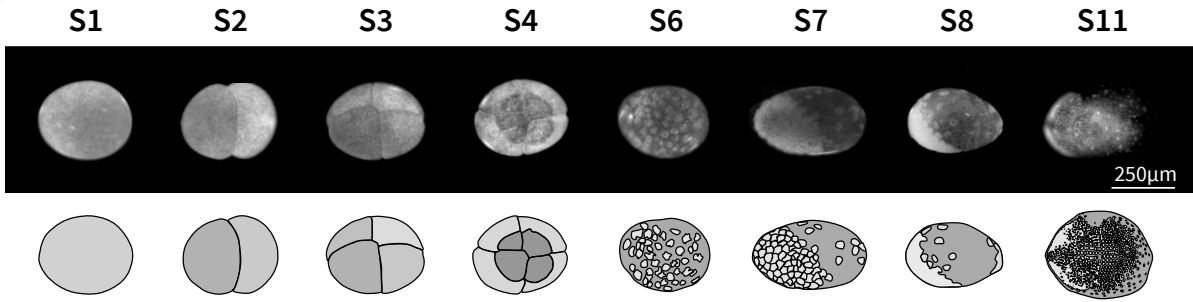
Figure 4.1 (following page): Life cycle and early embryogenesis of *P. hawaiiensis* adapted from ^{4,6,7,21,22} **a)** Life cycle of *P. hawaiiensis*, adult male and female form a mating pair (or a couple) resulting in the female laying eggs in a brood pouch. Those eggs will develop into a juvenile within 10 days. Sexual maturity is achieved within 3 months. hpf: hours post-fertilization, dpf: days post-fertilization, w: weeks, d: days. **b)** Above, white light microphotographs of *P. hawaiiensis* embryos during the early stages of embryogenesis from the one-cell stage (S11) to the elongation of the Germ Band (S11). Below, schematic representation of the stages shown above, those schematics are reused throughout the document. **c)** Schematic representation of the fate of the eight blastomeres of *P. hawaiiensis* during the early stages of embryogenesis. All germ layers are represented and specified by the 8 cell stage (S4). Adult *P. hawaiiensis* animals, male and female.

Figure 4.1: (continued)

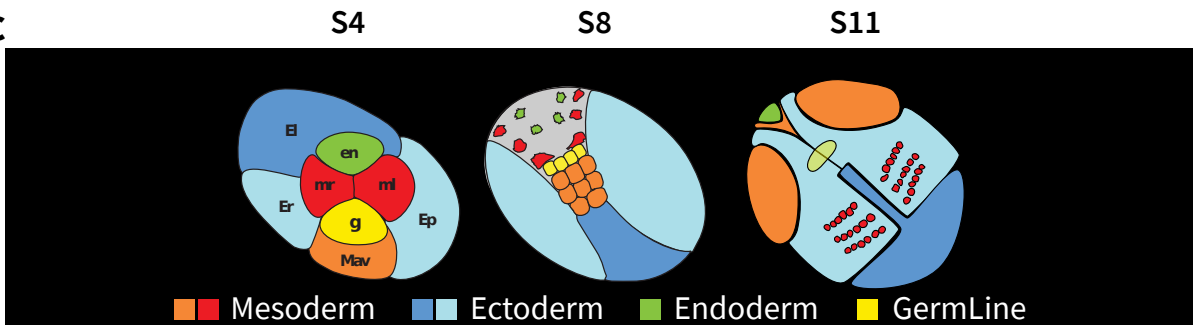
a



b



c



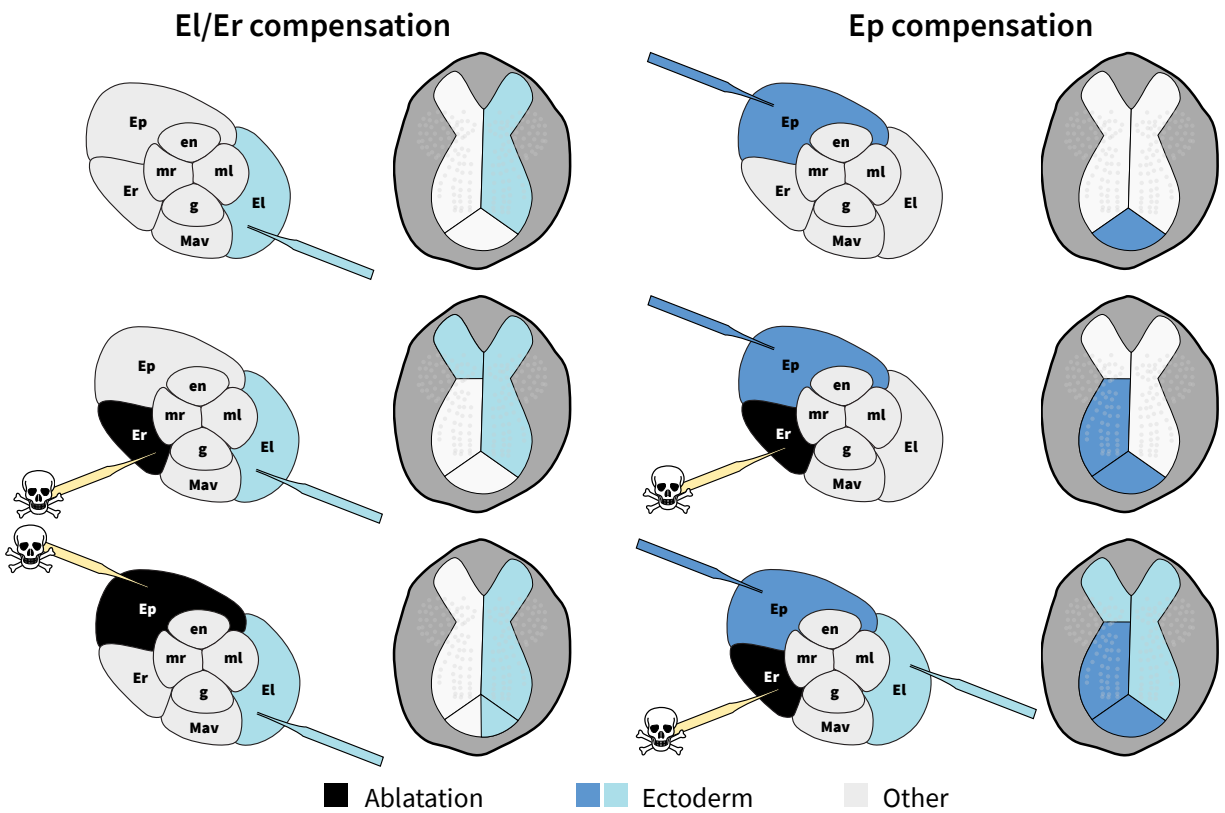


Figure 4.2: Schematic representation of the blastomere ablation and intra-germ layer compensation experiments for the ectodermal lineages adapted from¹⁵. Each of the three ectodermal blastomeres, El Er and Ep, were ablated while another blastomere was injected with a dye. The results of the experiments are presented in the schematic representation of germ band embryos.

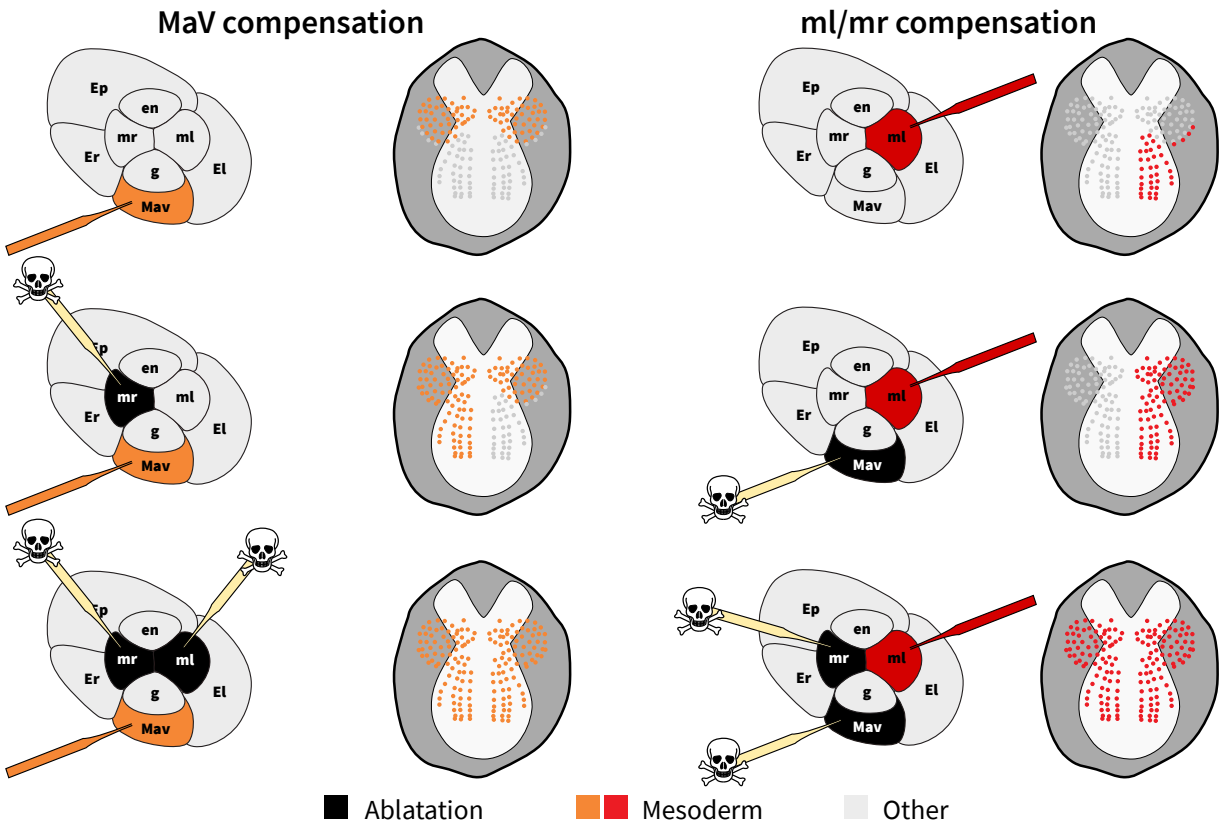


Figure 4.3: Schematic representation of the blastomere ablation and intra-germ layer compensation experiments for the mesodermal lineages adapted from ¹⁵. Each of the three mesodermal blastomeres, ml, mr and Mav, were ablated, while another blastomere was injected with a dye. The results of the experiments are presented in the schematic representation of germ band embryos.

4.1.2 The maternal to zygotic transition of *P. hawaiiensis*

One fascinating phenomenon during animal development is the activation of the zygotic genome. Indeed, in most metazoan embryos, the formation of the egg is accompanied by the maternal deposition of key mRNA transcripts (reviewed by Tadros and Lipshitz²³). Maternal transcripts play a key role in the first stages of embryonic development (reviewed by Tadros and Lipshitz²³), such as the gene *oskar* discussed in Chapters 0 to 2²⁴. However, while the correct expression and segregation of transcripts in early development are essential to initiate early differentiation (reviewed by Tadros and Lipshitz²³), the later activation of zygotic genetic programs is required to continue the differentiation process (reviewed by Tadros and Lipshitz²³). This transition is called the Maternal to Zygotic Transition (MZT) (reviewed by Tadros and Lipshitz²³). To study the changes in gene expression specifying the germ layers of *P. hawaiiensis* the MZT must happen prior to or during the starting time of the study. An earlier study that measured the activation of transcription via the phosphorylation of the Serine 2 of the RNA pol II C terminal domain of embryos as early as the S1 stage up to the end of the S5 stage (100 cells) had suggested that the MZT happens during the S5 stage (starting at 32 cells and finishing at 100 cells) (Figure 4.4)¹⁰. Moreover, the activation of the genome in *P. hawaiiensis* does not happen uniformly across all cells as previously shown by two populations of nuclei at the 32 cell phase of S5, one with detectable RNA pol II Ser2P and one without¹⁰. In this chapter, I will present results of experiments aimed at improving our understanding of the activation of transcriptional variability seen in *P. hawaiiensis*.

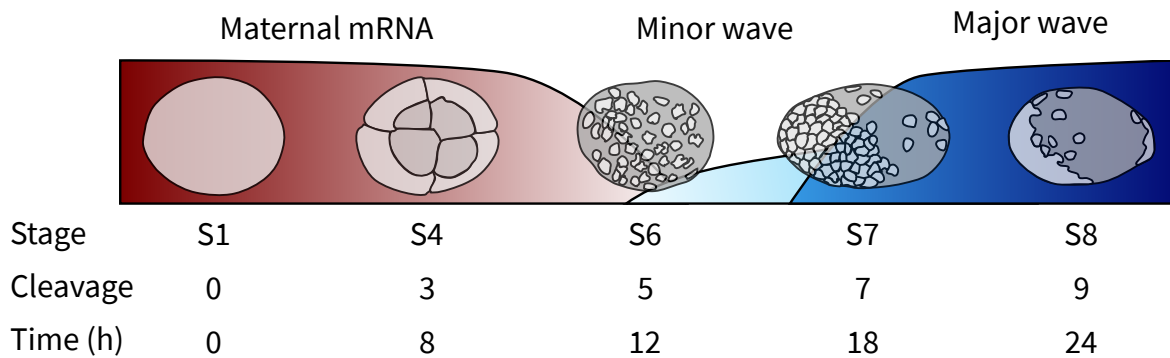


Figure 4.4: Schematic representation of the putative Maternal to Zygotic Transition of *P. hawaiiensis* adapted from the results of Nestorov et al.¹⁰ with the schematic representation from the review paper Tadros and Lipshitz²³. Speculative extrapolation of the timing of the activation of transcription from the beginning of the S5 stage (32 cells) with the minor wave, and general activation at the end of the S6 stage (128 cells) with the major wave.

4.1.3 The use of single-cell RNA sequencing to study embryonic development

In 1940, Conrad H. Waddington published a book called *Organisers and genes* in which he proposed a model called the "epigenetic landscape"²⁵. Under this model, cells possess a differentiation potential resting on a landscape. Through forces that he called "epigenetic", the landscape's shape was modified, inducing the cell to "move" through it to arrive at its final differentiated destination²⁵. Using recent advances in single-cell RNA sequencing, we can attempt to reconstruct the differentiation trajectories of cells in time by measuring their gene expression landscape. Recent reports of the study of the development of *C. elegans* with Cel-Seq²⁶, *Xenopus laevis* with inDrop²⁷, and *Danio rerio* with 10X^{28,29} are examples of such attempts to understand at the single cell resolution the mechanisms that determine the differentiation of cells.

These techniques all rely on the embryo being dissociated into single cells, thereby losing the spatial information necessary to fully understand development. In *D. rerio*, by using published expression patterns of genes to create a positioning system, sequenced single cells could be relocated to general embryonic areas³⁰. Other technologies such as single-molecule fluorescent *in situ* hybridization (smFISH) that allow for the accurate localization of transcripts could in principle be used to generate single-cell resolution maps of transcripts³¹ but suffer

from low throughput in term of mRNA diversity. In cultured cells, newer techniques such as MERFISH offer a higher throughput by using a fluorescent probe encoding system, but have not yet been successfully applied to complex tissues such as a developing embryo^{32,33}. One of the technical goals I had initially set for this chapter was to generate a new method that would allow for the successful geo-positioning of sequenced single cells onto a developing embryo at the single-cell resolution without the need for previously known landmark gene expression patterns. By studying the morphogenesis of a developing embryo (discussed in Chapter 5), we can observe all cells *in situ*, and know both their lineage and their positional relationships. By sequencing the transcriptome of the cells in the embryo we can determine the internal gene expression state of the cells. By combining visual and sequence data, I hoped to understand the differentiation processes and cell fate acquisitions in the early stages of development of *P. hawaiiensis*.

4.1.4 Mapping single-cell transcriptomes onto a virtual developing embryo to understand the differentiation of the germ layers

Both cell lineage identity (Figure 4.5 I)⁹ and cellular differentiation trajectories (Figure 4.5 II)^{27,29} can be modeled using a tree. While not directly related, both trees share a common feature, the developmental and morphogenic state of the embryo. One aim of both this chapter and Chapter 5 was to find a common feature between both trees such that the information they contained could be merged into a single developmental tree that would encompass lineage, position, dynamics, and gene expression information. I refer to this process as obtaining a geo-positioning system for single-cell RNA sequencing towards making a developmental atlas of embryogenesis. I believe that trying to merge both trees without any processing would be unlikely to result in accurate mapping. First, I reasoned that the time scale at which cells divide and cells differentiate (measurable changes in expression profiles) might not correlate, such that the branching events between both trees would not necessarily overlap. Second, the lineage tree is at the single-cell resolution while the differentiation tree is at the "cell type", or cell cluster, resolution. Due to this, the lineage tree might harbor a large number of branches while the differentiation tree might have fewer branches. To overcome those problems, I reasoned that I

would need to reproject both trees onto another tree that would serve as an abstraction to remap the information in both original trees (Figure 4.5 III). My proposed solution would be to enrich the branches of the lineage tree using features extracted from the microscopy images, including velocity, cell shape, cell size, and cell position relative to the major embryonic axes (Figure 4.5 II). Such features have been successfully used together to inform cell clusters and infer molecular mechanisms in the study of the Zebrafish posterior Lateral Line Primordium³⁴. I expected that I would have to scale the differentiation tree appropriately to match the projected lineage tree (Figure 4.5 III). Branching events could then be correlated to changes in cell clusters of the lineage tree. Finally, by combining all trees, I hoped that it would become possible, potentially at the single-cell resolution, to position gene expression profiles from the RNA sequencing dataset onto a 3D developing embryo (Figure 4.5 IV).

For this approach to be feasible, it would require accurate tracking of nuclei from as early in development as possible, as well as a low enough number of such cells to make this tracking feasible. Therefore, the organism *P. hawaiiensis* seemed perfectly suited for this procedure. However, a large number of common protocols in other model organisms such as *D. melanogaster* are not yet developed for this emerging model organism. The main challenge of this section of my Ph.D. was the large amount of protocol development that had to be done to achieve the goals set forth. Here and in Chapter 5, I present the advances that I made towards generating the datasets that would have been required for the geospatial mapping of single-cell RNA sequenced cells onto a developing embryo.

While the idea above directed the experimental procedure described below, most of this chapter is dedicated to the presentation of the developments of the protocols required to obtain single cell RNA sequencing data from embryos. Due to multiple unforeseen difficulties, no usable transcriptomic data was obtained. In this chapter, I describe the attempts and failures that led to the development of a single cell dissociation protocol for stage S6, S7, and S11 *P. hawaiiensis* embryos. Finally, I describe the resulting libraries from two inDrop encapsulation experiments and propose some rationale for the lack of depth in sequence data generated by those libraries.

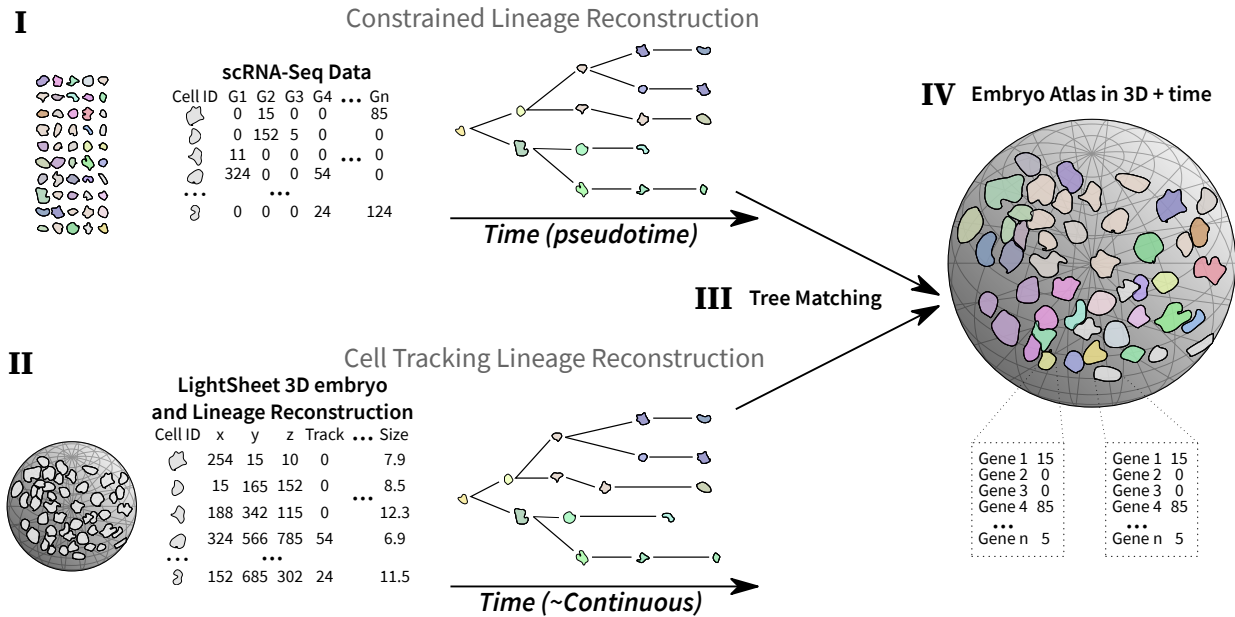


Figure 4.5: Schematic representation of the concept of the geo-positioning system presented in this chapter. In **I**), cells from embryos at different time points are sequenced and their differentiation tree is reconstructed. In parallel in **II**), embryos are imaged using light sheet microscopy, lineages are tracked and cell features are extracted. In **III**), the trees are matched against each other and the scRNAseq cells are mapped back onto the developing embryo. Finally, in **IV**), gene expression is projected onto a virtual developing embryo.

4.2 Methods

4.2.1 *P. hawaiiensis* cultures

Three cultures of *P. hawaiiensis* were initiated from animals kindly provided by Anastasios Pavlopoulos (Institute of Molecular Biology and Biotechnology, Greece): The original line from the Chicago aquarium (referred to as Wild Type)⁵, an Isogenic female line created by Anastasios Pavlopoulos from which the genome was derived⁴ and a genetically modified line containing an *H2B-mRFPruby* cassette under the control of a heat-shock inducible promoter (*PhHS>H2B-mRFPruby*)⁹. The three lines were put in culture in artificial seawater (ASW) (Instant Ocean Artificial Seawater #671442) with a specific gravity of 1.020 (*P. hawaiiensis* is resilient to salt concentration and will develop with a specific gravity ranging from 1.018 to 1.022). Each tank consisted of a bottom layer of two centimeters of sterilized crushed coral and an aquarium bubbler fed by an aquarium air pump. The crushed coral was sterilized by being boiled in distilled water for 2h. Each tank received a 50-75% water change every 7-10 days depending on

the concentration of animals in culture (the higher the concentration the more frequent the water change).

Animals were fed a mixture of $\frac{1}{3}$ Tetra Tetramin Flakes, $\frac{1}{3}$ wheat germ, $\frac{1}{3}$ Tetra Algae Wafers. The food preparation was done by mixing the three ingredients in a clean bowl and grinding them with a mortar into a fine powder. The powder was then used directly by sprinkling it over the water in the culture tank twice a week. However, this tends to increase the turbidity of the water quickly, therefore a second approach was later used, as follows: a layer of powdered food 6-7 mm thick was laid into a 15cm petri dish. Then 50ml of liquid 1% agar in sterile distilled water was poured into the dish. The food and the agar were then mixed thoroughly and left to rest in a 4°C fridge until the agar had solidified. This preparation can be kept at 4°C for 2 weeks at most. The gel is cut into 2cmx2cm pieces, and each cube was then dropped into the tank. One to two cubes were fed to a large tank every week.

To avoid loss of an entire line, clones of the cultures were kept at all times, such that in the event of a total population collapse in one tank, that line would not be lost. Moreover, all tanks had two air pumps feeding into the bubbler to guard against the malfunction of a given pump.

4.2.2 Collection of *P. hawaiiensis* embryos

P. hawaiiensis males will attach to a fecund female with the use of the front gnathopods, resulting in a pair of individuals that are called a couple. Once the female has received a spermatophore from the male, it will molt, freeing itself from the male, and will shortly after lay her eggs in an external ventral brooding pouch. Embryos can then be harvested from the single females for subsequent use.

Couples are easily visible by eye in the culture tank and can be captured with a small net. To create a capturing net, a 50ml falcon tube and a 200um nylon mesh are combined as shown in Figure 4.6:

Once couples have been captured in the net, they are placed in a petri dish filled with ASW for subsequent use.

Couples were harvested the evening before an experiment and placed in Petri dishes. On the

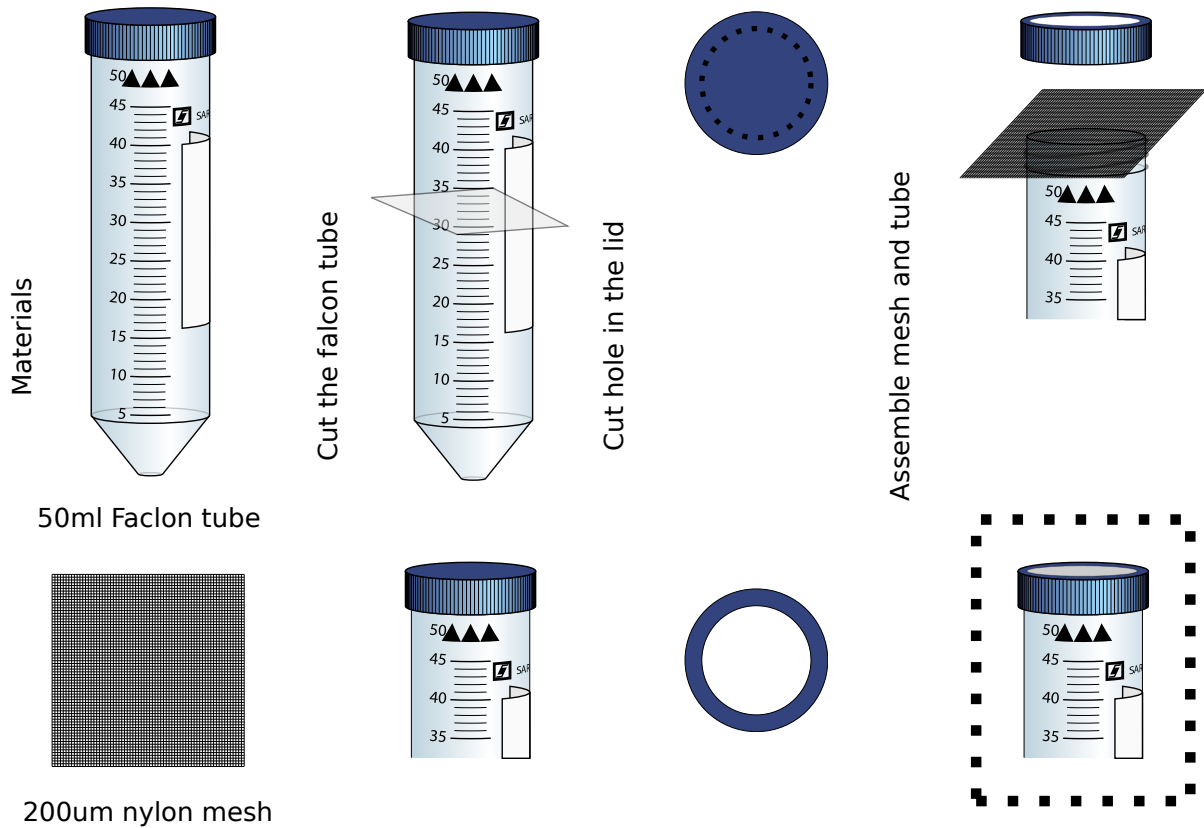


Figure 4.6: Schematic representation of the creation of a mesh trap for the capture of *P. hawaiiensis* adults. 50ml Falcon tube and 200um mesh are used to create this net. Cut the falcon tube between the 30 and 35ml mark using a metal saw or a hot wire. Cut a circular hole in the lid which will allow for the water to flow out of the net. Cut a 200um nylon mesh into a square piece of side length the diameter of the falcon tube plus 5mm. Place the mesh over the screw part of the tube and screw the lid back, trying to keep the mesh flat and under tension.

morning of the experiment, detached females were collected from the Petri dishes, and anesthetized using ASW saturated with CO₂. To saturate the ASW, a stone bubbler connected to a CO₂ tank was placed in a bottle of ASW for 3 minutes. Most of the ASW in the Petri dishes with females was removed and replaced by the CO₂-saturated ASW. In this medium, the females will go unconscious within one minute or less. Using a dissecting microscope and forceps, the embryos were removed from the brooding pouch. Embryos were checked to determine their developmental stages, and one cell stage embryos were kept in filtered artificial seawater with antibiotics (FASWA) (3ml of pen/strep (10000U/ml Thermo: 15140122) + 1.5ml amphotericin (250ug/ml stock 15290-026 Gibco life technology) in 300ml of filtered artificial seawater (FASW)) for subsequent injection and imaging. This protocol was derived from standard practices published in Kontarakis and Pavlopoulos³⁵.

4.2.3 Heat Shock of PhHS>H2B-mRFPruby *P. hawaiiensis* embryos

P. hawaiiensis embryos were collected as described above (see: *Collection of P. hawaiiensis embryos*), then placed in a 5cm petri dish filled with FASW. The embryos were placed in an incubator set at 37°C for 60 minutes, then were left at room temperature for 30 minutes. After the heat shock, the FASW was replaced with an equal volume of fresh FASW to avoid the effect of increased salinity due to water evaporation during the heat shock period. Embryos were then kept in a 26°C incubator. All embryos were imaged within 12-24h after the heat shock was performed.

4.2.4 Imaging of *P. hawaiiensis* embryos with the Zeiss AxioZoom fluorescent stereomicroscope

P. hawaiiensis embryos were placed in a 5cm or 10cm petri dish filled with FASW. The petri dish was placed under the Zeiss AxioZoom microscope on an air table to avoid vibrations and specimen drift. Images were taken using a Hamamatsu OrcaFlash4.0 v2 camera installed on the microscope. Illumination for white light imaging was provided from a white light source through two focusable optic fiber connections, resulting in an incident light illumination. The angle of incidence was adjusted each time to provide the best visible contrast.

4.2.5 Dissection and fixation of *P. hawaiiensis* embryos

P. hawaiiensis embryos were collected as described above (see: *Collection of P. hawaiiensis embryos*), then placed in a 5cm petri dish filled with FASW. Stages were selected using a dissecting microscope with incident lighting. The fixation and dissection procedure was followed exactly as previously described in *Fixation and Dissection of P. hawaiiensis Embryos*³⁶. Embryos were kept in 100% Methanol at -20°C for later use or used directly after post-fixation washing for immunostaining.

4.2.6 Collection and fixation of *D. melanogaster* embryos

D. melanogaster adults were put in a collection cage with a 15cm apple juice/agar petri dish³⁷ and left at 25°C for four hours before collection. The 15 cm petri dish was removed from the cage and replaced with a ten cm apple juice petri dish with 1g of wet yeast³⁸. After ten minutes, the petri dish was removed and embryos were collected as previously described in Rothwell and Sullivan³⁸. The embryos were dechorionated with 50% bleach for 1min and fixed with a 50/50 mixture of heptane/4% Paraformaldehyde (PFA). They were then dehydrated in methanol and either stored at -20°C or rehydrated and used directly for staining as described below.

4.2.7 Immunostaining of RNA Polymerase II in *P. hawaiiensis* embryos

Fixed embryos were stained using the protocol *Antibody Staining of P. hawaiiensis Embryos*³⁹, with the following modifications: Because the secondary antibodies used were conjugated to fluorescent proteins no chemical development was performed on the embryos. The washing steps were followed by an additional overnight wash in 1X PBT at room temperature. After the last washing step of the secondary antibody, embryos were mounted as described in *Confocal imaging and mounting of P. hawaiiensis embryos* and *Preparation and analysis of fixed embryos for light sheet microscopy with the Zeiss Z1 microscope* for subsequent imaging. Three primary antibodies were used against RNA Polymerase II C-terminal domain (CTD) at 1:500: Purified mouse anti-RNA Polymerase II RPB1 Antibody H5 (BioLegend 920203) which targets the CTD when the Serine 2 is phosphorylated, Purified mouse anti-RNA Polymerase II RPB1 Antibody

H14 1:500 (BioLegend 920304) which targets the CTD when the Serine 5 is phosphorylated and Go-ChIP-Grade™ Purified mouse anti-RNA Polymerase II RPB1 Antibody Clone 8WG16 at 1:500 (BioLegend 664911) which targets the unphosphorylated CTD. One primary antibody targeting Tubulin was used as a positive control: Mouse anti-alpha Tubulin antibody [DM1A] - Microtubule Marker at 1:1000 (Alexa Fluor® 488) (Abcam ab195887). For *D. melanogaster* embryos, a rabbit anti-Vasa antibody was used to mark pole cells at 1:100 (Vasa d-260). All secondary antibodies were used at 1:1000. Donkey anti-mouse conjugated with Alexa fluorophore 488 and 555 were used as secondary antibodies: Donkey anti-Mouse IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 488 (Thermo Fisher #A-21202 RRID: AB_141607) and Donkey anti-Mouse IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 555 (Thermo Fisher #A-31570 RRID: AB_2536180). For Vasa a goat anti-Rabbit conjugated with the far red Alexa 647 dye was used: Goat anti-Rabbit IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor Plus 647 (Thermo Fisher #A-21246 RRID: AB_2633282).

Pole cells in wild type *D. melanogaster* embryos at 75-85 minute post-fertilization undergo the MZT with a delay compared to the somatic nuclei⁴⁰, providing an internal control. In all experiments, *D. melanogaster* embryos at 75-85 minutes post-fertilization were added to the tube containing fixed *P. hawaiiensis* embryos, and stained and imaged using the same confocal settings as the *P. hawaiiensis* embryos.

4.2.8 Confocal imaging and mounting of *P. hawaiiensis* embryos

P. hawaiiensis embryos are ~500um diameter and spherical. Embryos were mounted between a glass slide and coverslip with wax feet. First, a single embryo was deposited on a glass slide and the excess liquid removed, then a drop of Vectashield (VectorLabs: H-1000) supplemented with a 1:10000 Hoechst 33242 dilution was deposited onto the embryo. The coverslip was prepared by scraping each corner against Dental wax (Amazon: B00FKDCoUU) until a small amount (~1mm) of wax accumulated. Then, the coverslip with wax risers was deposited atop the drop of vectashield and embryo. Finally, the coverslip was gently secured by pushing with forceps on each corner until the embryo was just touching both the slide and the coverslip. The procedure

is represented in Figure 4.7.

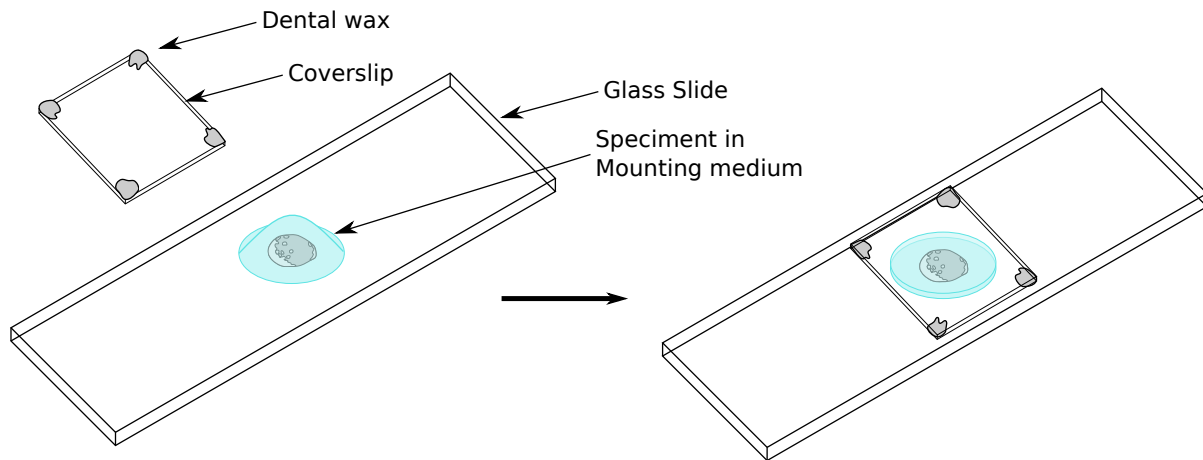


Figure 4.7: Schematic representation of the mounting procedure on a glass slide and coverslip for *P. hawaiiensis* embryos. The coverslip is padded with dental wax. The embryo is mounted in Vectashield. The coverslip is placed gently atop the Vectashield drop and fixed by gently pressing each corner.

4.2.9 Preparation and analysis of fixed embryos for light sheet microscopy with the Zeiss Z1 microscope

Embryos were mounted in 1% low melt agar in PBS with fluorescent beads (see: Chapter 5 *Preparation of low melt agar gel with fluorescent beads*) inside a glass capillary affixed with a micro Teflon plunger. Glass capillaries (Brand #708718) and plungers (Brand #701932) were provided by Seth Donoughue. To mount *P. hawaiiensis* embryos, the black color-coded capillaries (BlauBrand: 708718) and plunger were used. Embryos were dropped into liquid 1% low melt agar with fluorescent beads kept at 42°C and stirred until they dropped to the bottom of the tube. Then using the capillary and the plunger, first ~2cm of agar was aspirated, then the embryos, one at a time to avoid having them mounted at the same height in the tube. Finally, 2-3mm more agar was added. The tips of the capillary were kept submerged in PBS until the samples were mounted in the Zeiss Z1 microscope.

Each embryo was annotated with Mamut, a FIJI plugin for light sheet dataset cell tracking, and every nucleus was annotated. Then, from this set of annotations, nuclei that did not show a signal for RNA Pol II CTD Ser2P or Ser5P were removed. Both annotation sets were then imported into python for subsequent analysis (Figure 4.11).

4.2.10 Isolation of *P. hawaiiensis* embryo blastomeres

Embryos were collected as described previously (see: *Collection of P. hawaiiensis embryos*) and staged to select for four cell stage embryos and placed in FASWA. The isolation of single blastomeres was performed as previously described by Cassandra Extavour⁴¹.

4.2.11 Dissociation of *P. hawaiiensis* embryo into single cells

The dissociation of *P. hawaiiensis* embryos was carried out as described in the results section:

Detailed protocol for *P. hawaiiensis* embryo single-cell dissociation. Over 100 couples were collected into 15cm Petri dishes filled with ASW. After one hour, detached females were removed from the Petri dishes. Couples were left at 26°C for the following four hours, then detached females were harvested and placed in a separate 15cm petri dish filled with ASW. This was performed such that the exact timing of embryonic development could be assessed up to a four hour (or shorter if needed) time window. Embryos were then collected from anesthetized females (see: *Collection of P. hawaiiensis embryos*) and placed in FASWA prior to dissociation. All material used for the cell dissociation was coated with Bovine Serum Albumin (BSA) by placing a 1% BSA in sterile water solution in the Petri dishes, pipettes tips, and glass Pasteur pipettes. The solution was left to passively coat the material for 12 hours. The material was then quickly rinsed in sterile water before being used for the dissociation. The embryos were first washed two times with the dissociation buffer composed of Isethionic acid 15mg/ml (Isethionic acid sodium salt: Sigma-Aldrich 220078-25G), Sodium pyrophosphate 9mg/ml (Sodium pyrophosphate tetrabasic decahydrate Sigma-Aldrich S6422-100G), and CAPS 2.2mg/ml (CAPS Sigma-Aldrich C2632-25G) to remove any excess calcium from the solution. Embryos were then dissected out of the eggshell as gently as possible while minimizing any damage to the cells. The dissection was performed on a BSA-coated Sylgard plate using tungsten needles in batches of 30 embryos (working with batches of more embryos resulted in decreased dissociation efficiency) in a drop of dissociation buffer. The dissected embryos were transferred to one well of a BSA coated 12 well plate using a BSA coated glass Pasteur pipette. Extra care was taken not to transfer any of the eggshells as they interfere with the dissociation. The well was filled entirely with dissociation buffer, then covered with parafilm such that no air bubble got trapped in the

liquid column. The plate was then sealed with tape and vortexed for 25 minutes. The plate was removed and cells were then transferred onto a 15ml BSA coated tube. The next steps are detailed in Figure 4.22. At the bottom of the tube, using a 21G needle, a gradient density of optiprep was backfilled. First, 60ul of 5% Optiprep in 1x PBS with Phenol Red dye (OptiPrep™ Density Gradient Medium, Sigma: D1556), followed by 60ul of 10% Optiprep in PBS, then 20ul of 20% Optiprep in PBS, and 20ul of 30% Optiprep in PBS with Phenol Red dye and finally 50ul of 40% Optiprep in PBS. The cells were concentrated by centrifuging the tube at 2500rpm in a bucket centrifuge for 4 minutes at 4°C. All liquid above the first red band (5% Optiprep) was removed by gentle pipetting, then, using a 1ml syringe, the content between the two red bands was harvested, making sure not to aspirate the 40% optiprep containing debris. This solution of dissociated cells was then kept at 4°C and used for subsequent experiments.

4.2.12 Assessment of dissociation efficiency, cell concentration, and cell viability

Dissociation efficiency was assessed using a hemocytometer. 20ul of the cell suspension was added between the hemocytometer and the coverslip. Cells were counted using differential interferometry contrast (DIC) white light imaging on a Zeiss AxioScan. Every single cell, doublet, triplet, or aggregate of more than three cells was counted. The dissociation efficiency was calculated by taking the ratio of single cells to the total number of counted cells.

Cell viability was assessed using the Propidium Iodide (Thermo: P3566) staining method. Dissociated cells were incubated with a 1:5000 Hoechst 33242 (stock at 1 mg/mL) and 1:5000 PI (stock at 1 mg/mL) for 1 minute. 20ul of the cell solution was then placed in a hemocytometer to quantify cell concentration. Finally, all Hoechst-positive and PI-positive cells were counted using an epifluorescent upright microscope (Zeiss AxioScan). The viability was calculated by taking 1 minus the ratio of PI-positive cells to the total number of cells.

4.2.13 Preparation of CelSeq2 libraries

Embryos earlier than stage S6 were dissected out of their eggshells with tungsten needles, and cells were separated using an eyelash as previously described⁴¹. Individual cells from dissociated embryos were pipetted on to the lid of a LoBind PCR 96 well plate under a dissection microscope. The lids were then placed on the corresponding PCR 96 well plates and the plate stored at -80°C. The library preparation protocol was then applied, adapted from the official CelSeq2 protocol from the original paper⁴². The only difference was the use of the BioMek pipetting robot to perform all the pipetting after the first primer annealing.

4.2.14 Preparation of inDrop libraries

Embryos were harvested and staged such that Stage 6-7 (12-24hpf) and Stage 11 (64-68hpf) were dissected and dissociated as described above (see: Dissociation of *P. hawaiiensis* embryo into single cells). 160 embryos were used per experiment, with two replicates per time point. inDrop libraries were prepared by the Harvard Medical School Single Cell Core (<https://singlecellcore.hms.harvard.edu/>)⁴³. The libraries were sequenced on the NextSeq platform at the Harvard Bauer Core Facility (<https://bauercore.fas.harvard.edu/>) with an estimated 100,000 reads per cell. Libraries were sequenced Paired-end with the special inDrop protocol (protocol developed by the Harvard Bauer Core Facility, personal communication with C.B. Reardon) which produced 61bp reads for the 3' end of the captured transcripts, along with the 3 required barcodes: Library, Cell, and UMI.

4.2.15 Annotation of *P. hawaiiensis* genome

The version of the *P. hawaiiensis* genome, v5.0 (GCA_001587735.2 Phaw_5.0), was annotated using the MAKER2 annotation tool suite⁴⁴. Unpublished RNA sequencing reads for use in annotation were kindly donated from the laboratories of Michaelis Averof (Institut de Génomique Fonctionnelle de Lyon, France) , Nipam Patel (Marine Biological Laboratory, Woodshole USA), Ezio Rosato (University of Leicester, Leicester, UK), Anastasios Pavlopoulos (Institute of Molecular Biology and Biotechnology, Greece) and the Extavour lab. All published RNA

sequencing libraries were collected from published databases as well (NCBI IDs: PRJNA399131, PRJEB2845, PRJEB2844). Transcript data from other crustacean species were added to the maker pipeline as external gene evidence (NCBI TSA ids: *Gammarus pulex* HAFM01, *Hyalella azteca* GAJP01 GAJQ01 GEHV01 JQDR02, and *Talitrus saltator* GDUJ01).

First, the detection of repeated regions and transposons was generated to create a mask for further filtering. A pipeline developed by Guillem Ylla in the Extavour lab was used. It consisted of running the following repeated region detection algorithms and matches against databases: MITE-tracker⁴⁵, LTRharvest + LTRDigest⁴⁶, RepeatModeler⁴⁷, SINE database⁴⁸, RepeatMasker⁴⁹ and RepBase⁵⁰. The algorithms were run with default parameters. The results were then joined into a single file and regions classified using RepeatClassifier⁴⁷. Possible protein-coding genes detected as transposable elements were filtered out of the set. Elements classified as "Unknown" by the RepeatClassifier and with a Blastx hit (e-value < 1e-10) against the insect proteins from the well-defined SwissProt database were removed. Finally, the genome was masked using RepeatMasker⁴⁹. This mask was used as a repeat mask in the MAKER2 pipeline.

Second, all reads from the RNA sequencing datasets were mapped against the genome sequence to create transcript models. Reads were mapped onto the genome using hisat2⁵¹ with default parameters. The SAM files outputs were converted to BAM files, concatenated, and sorted using SAMtools⁵². The BAM read maps were then put into the gene model prediction algorithm StringTie⁵³. Finally, the gene model was converted to GFF3 format using the Cufflinks built-in converter⁵⁴. This GFF3 file was used as gene evidence in the MAKER2 pipeline.

Third, a *de novo* gene prediction algorithm model was constructed for Augustus⁵⁵ using the BUSCO arthropod dataset as input⁵⁶. BUSCO looks for highly conserved genes in the *P. hawaiiensis* genome and automatically launches the Augustus training pipeline to fit a gene model. This gene model was used as the *de novo* Augustus model in the MAKER2 pipeline.

The MAKER2 pipeline was then executed on the Harvard High Processing Cluster (HPC) Odyssey with the following parameters:

- EST: The assembled transcripts for *P. hawaiiensis* published previously⁴
- ALTEST: The RNA sequencing transcripts collected from NCBI mentioned above

- EST_GFF: The GFF3 file generated above
- PROTEIN: The 2018 Uniprot Swiss Prot database⁵⁷
- RMLIB: The repeat masker output generated above.
- SOFTMASK: Turned on
- AUGUSTUS_SPECIES: The Augustus model generated above
- EST2GENOME: turned on
- PROTEIN2GENOME: turned on
- TRNA: turned on
- MAX_DNA_LENGTH: 300000
- AED_THRESHOLD: 1
- All other parameters were set to default.

The MAKER2 pipeline was run two consecutive times, with the second one using the generated annotations from the first. BUSCO⁵⁶ was used to compare the annotations against the *BUSCO arthropod database* to assess their quality.

4.2.16 Computational processing of inDrop libraries

The processing of the inDrop libraries was performed using a published demultiplexing pipeline⁴³. The pipeline was downloaded from the GitHub repository <https://github.com/indrops/indrops> and used with the default parameters except for the "d" parameter, where the maximum allowed distance between a hit and an annotation was set to 5000bp.

The analysis of the demultiplexed counts was performed using Scanpy, a single-cell RNA seq analysis toolkit in python⁵⁸.

4.3 Results

4.3.1 Attempts to examine the timing of *P. hawaiiensis* Maternal to Zygotic Transition

As we have seen above, the Maternal to Zygotic Transition (MZT) plays a central role in the specification of cell fate (reviewed in²³). A previous study had suggested that the MZT of *P. hawaiiensis* might start as early as the S5 stage (at 32 cells), based on the report that some cells at this stage, but not all, contained levels of RNA Polymerase II that were detectable with an antibody against a phosphorylated Serine diagnostic of polymerase activity¹⁰. The RNA polymerase II C terminal domain (CTD) is phosphorylated on Serine 2 prior to transcript elongation, and phosphorylated on Serine 5 when the polymerase is primed at the starting site (reviewed by Bartkowiak et al.⁵⁹. However, only the general timing of detection of this signal was reported¹⁰, namely that by the end of the S5 stage (reported as 100 cell stage), all nuclei displayed a positive signal for RNA Pol II Ser2P-CTD¹⁰.

To assess the timing of the MZT I first proposed to use an alternative method. By activating the expression of an inducible promoter controlling a reporter gene in a transgenic animal, I hoped to detect the timing of the MZT. I hypothesized that the promoter would only become inducible after the MZT happened. To try to refine the timing of the putative MTZ in *P. hawaiiensis*, we obtained a transgenic *P. hawaiiensis* line⁶⁰ containing a cassette expressing *H2B-mRFP* under the control of a heat-shock promoter *PhHS* (referred to as *PhHS>H2B-mRFP*) from Anastasios Pavlopoulos⁹. I first tested the transgene by subjecting S6-7 embryo (after the reported onset of the MZT during the S5 stage¹⁰) to a heat shock at 37°C s, and detected subsequent red fluorescence in the nucleus of all cells under an epifluorescence stereomicroscope (100% of embryos examined showed expression n=16) (Figure 4.8). This suggested that the transgenic cassette was expressing mRFP tagged H2B as expected.

I then tried a similar experiment with embryos at the S1-S5 stage (1 to 128 cell), I did not detect any transgene expression 12 hours following heat shock for any of them (Figure 4.8; sample size in legend). Given the previous report suggesting that the MZT might start as early as the 32 cell

phase of the S5 stage and continue through to the end of the S5 stage (100 cells)¹⁰, I was surprised that I could not detect any expression of the transgene in response to heat shock in 128 cell stage embryos. Therefore, I used the same transcriptional activity assessment technique as the previous report¹⁰, namely fixing embryos at different stages and detecting RNA Pol II CTD Serine 2 and Serine 5 phosphorylation states via immunohistochemistry. At stage S4, no positive staining was detected. The intermediate stage between the 16 and 32 cell phase of the S5 stage showed mosaic signal of phosphorylation of Serine 2 and Serine 5 of the CTD (Figure 4.9; n=6 for Ser5P CTD 4/6 showed a positive signal and n=3 for Ser2P CTD 1/3 showed a positive signal). By the 32 to 64 and 64 to 128 cell phases of the S5 stage, cells showing detectable anti-phospho-Serine 2 and -Serine 5 of the CTD were detected in all embryos (Figure 4.9; n=11 for Ser5P CTD and n=4 for Ser2P CTD, all embryos showed a positive signal in at least one nuclei). In all cases (n=6 for all conditions), *D. melanogaster* pole cells showed a negative Ser2P CTD staining and the somatic cells positive staining (Figure 4.10). To conclude, I found a mosaic expression pattern similar to the previously reported results¹⁰ but with an earlier onset at the transition between the 16 to the 32 cell phase of the S5 stage (compared to the reported 32 cell phase of the S5 stage).

One of the open questions that I hoped to answer was whether the timing of mosaic activation of RNA pol II was happening randomly or was restricted to particular lineages. Due to *P. hawaiiensis* early embryogenesis following a stereotypical division pattern^{1,7}, I hypothesized that by collecting a sufficient number of embryos, carefully dissecting them to preserve their 3D structure, and imaging them in 3D using a light sheet microscope, I could spatially align the images to assess whether the activation of certain cells as early as 16-32 cells during the S5 stage was restricted to specific lineages. Working with Robbert Appleby, an undergrad summer student from Cambridge University, we repeated the staining protocol as described above and imaged all embryos using a light sheet microscope. The collection of embryos, staining, and imaging were successful and 80 embryos across all conditions were imaged (Figure 4.11b-d, breakdown of sample size per stage in Figure 4.11a, n=37 for Ser2P, n=26 for Ser5P, n=17 controls). Due to the internship of Robert Appleby finishing before the completion of the annotations and to advance the second part of this chapter, namely the single-cell RNA

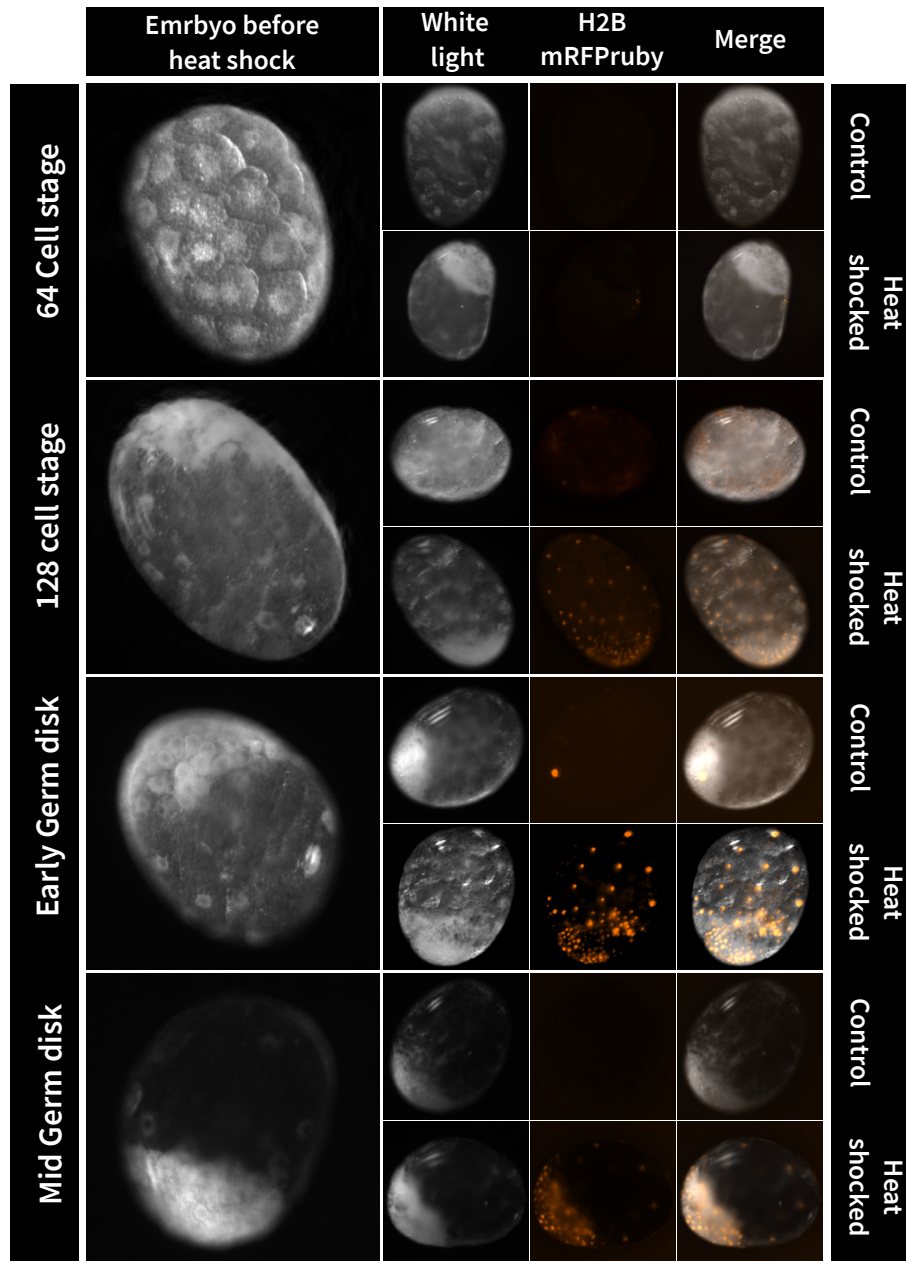


Figure 4.8: Representative images of heat-shocked embryos and control embryos. Shown are PhHS>H2B-mRFPruby transgenic embryos at four different developmental stages. For each experiment, a clutch of eggs from a single female was used and split in two, one set of eggs for control, and one set for heat shock treatment. S5 stage (64-cell): n=9, 4 controls, and 5 heat-shocked embryos; S6 stage (128-cell): n=12, 6 controls and 6 heat-shocked embryos; S7 stage (Early Germ disk): n=7, 3 controls, and 4 heat-shocked embryos; S7 stage (Mid germ disk): n=14, 7 controls and 7 heat-shocked embryos. Imaging was done on a Zeiss AxioZoom stereo microscope with an Orca Flash 4.0 camera.

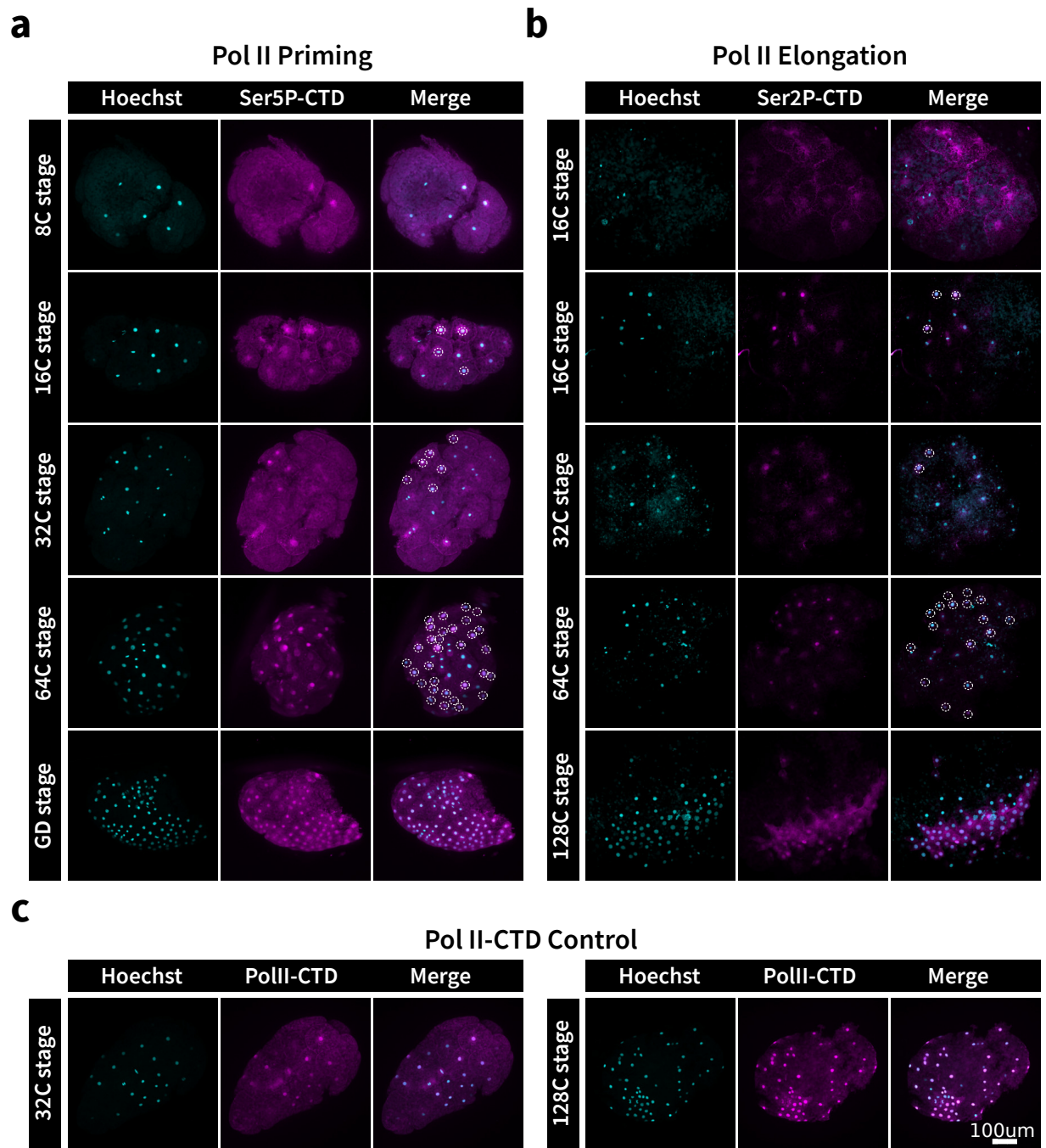


Figure 4.9: Representative images of *P. hawaiensis* embryos stained with antibodies against different phosphorylation states of the RNA Polymerase II CTD. a, b, c) Staining of *P. hawaiensis* embryos at different cell count of the S5 stage (16-128) and early S6 stage (GD stage) with Hoechst-33242 and monoclonal antibodies against the CTD. a) Staining with monoclonal mouse antibody clone H14 against Ser5P CTD, an indication of primed RNA Pol II. n=25 b) Staining with monoclonal mouse antibody clone H5 against Ser2P CTD, an indication of elongating RNA Pol II. n=10 c) Positive control staining with monoclonal mouse antibody clone 8WG16 against the unphosphorylated CTD. n=4

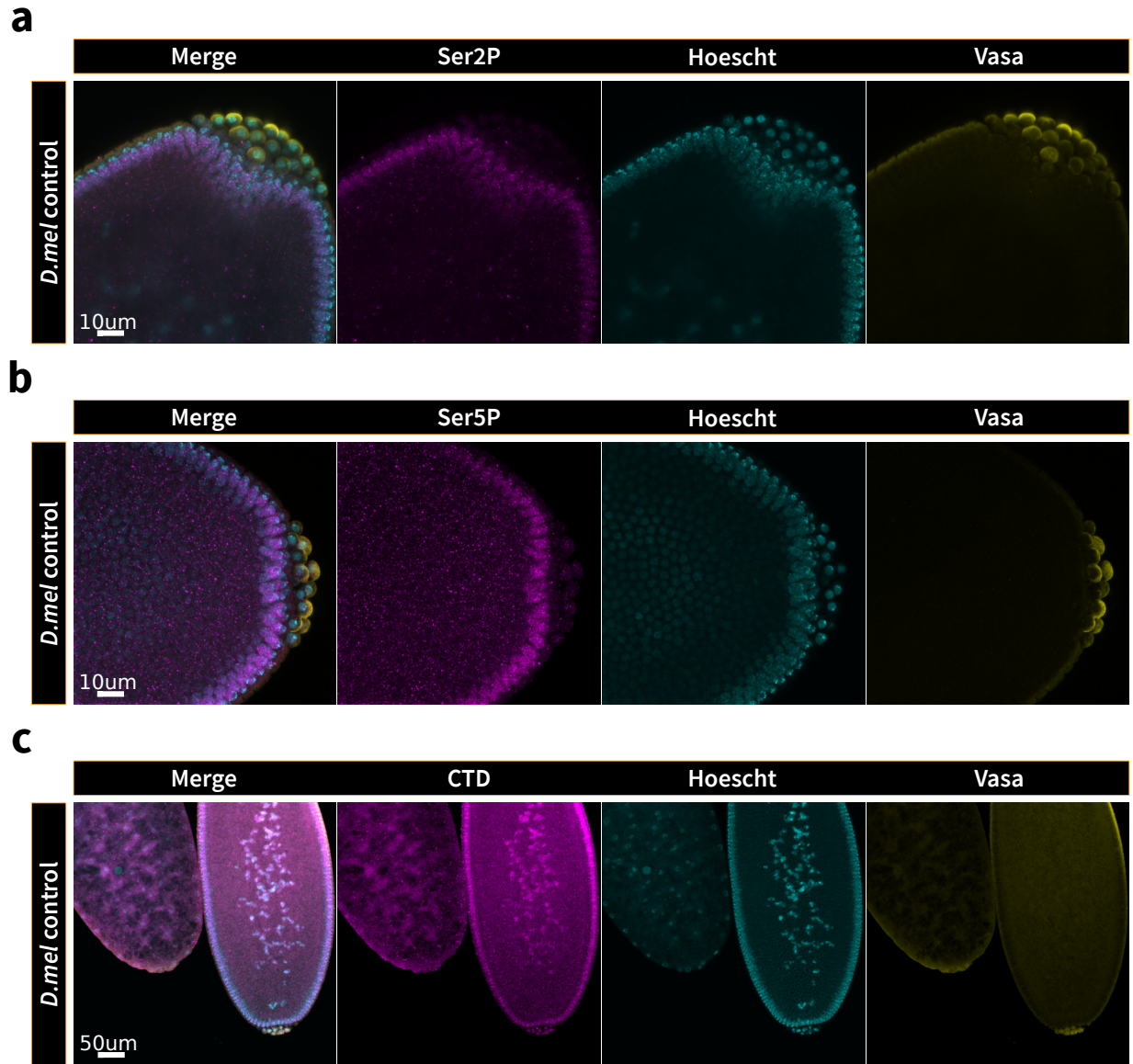


Figure 4.10: Positive control for the antibody staining presented in Figure 4.9. Representative images of *D. melanogaster* embryos stained with antibodies against different phosphorylation states of the RNA Polymerase II CTD and Vasa as a marker of pole cells. **a, b, c)** Staining of *D. melanogaster* embryos at 75-85mn post-fertilization (pole cell formation) with Hoechst-33242, anti-Vasa, and monoclonal antibodies against the CTD. **a)** Staining with monoclonal mouse antibody clone H5 against Ser2P CTD showing elongating RNA Pol II. n=6 **b)** Staining with monoclonal mouse antibody clone H14 against Ser5P CTD showing primed RNA Pol II. n=6 **c)** Positive control staining with monoclonal mouse antibody clone 8WG16 against the CTD. n=6

sequencing of embryos, I left the analysis of this dataset unfinished. The annotation of nuclei was partially completed and shows promising results (Figure 4.11e) whereby the partial activation of nuclei can be seen before the end of the S5 stage (128 cells) embryo (n=2 analyzed for Ser2P and n=1 analyzed for Ser5P).

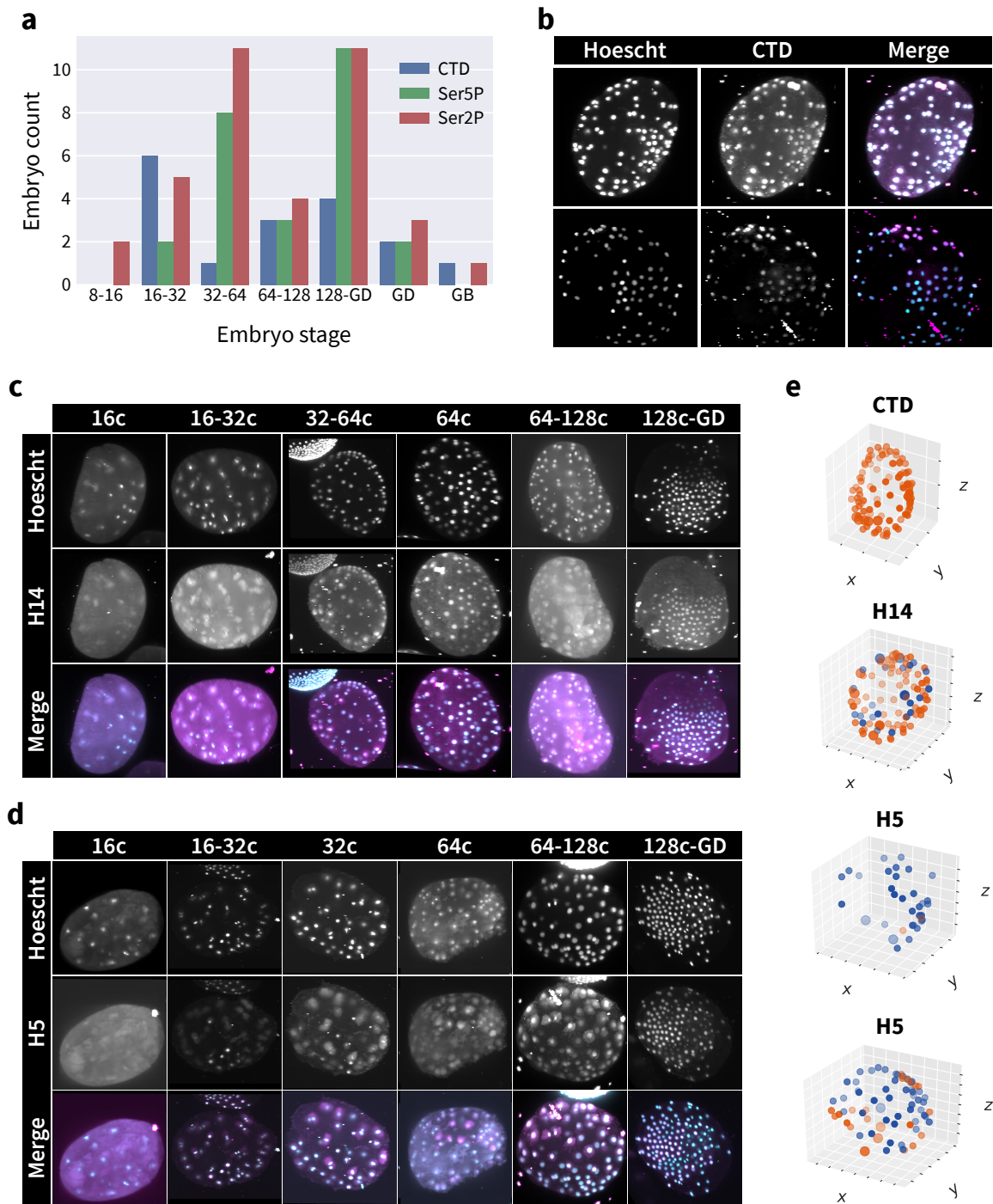


Figure 4.11: Light sheet imaging of *P. hawaiiensis* embryos stained with antibodies against different phosphorylation states of the RNA Polymerase II CTD. a) Distribution of stained embryos per antibody and stages. 69/80 of the embryos sampled to date are between the 16-128c stages. **b-d)** All images are maximum intensity projections of the entire embryo volume acquired with the Zeiss light sheet Z1. **b)** Representative image of a positive control 64-128c stage embryo stained for PolIII-CTD. **c)** Staged embryos stained with antibody H14 (PolIII CTD:Ser5P) showing primed PolIII. **d)** Staged embryos stained with antibody H5 (PolIII CTD:Ser5P) showing elongating PolIII. **e)** Example visualization of annotations of nuclei positions and transcription state. 3D scatter plot showing the position of the nuclei and its activation state for each antibody. In blue, the nuclei are negative for the stain, and in orange they are positive. From top to bottom, phases of stage S5 are 64-128c, 64-128c, 16-32c, and 64-128c.

4.3.2 Isolation of single cells for plate-based single-cell RNA sequencing and failure to generate single-cell libraries using the CelSeq2 protocol

Given the putative relatively early activation of the zygotic genome (Figure 4.9, Figure 4.11) in *P. hawaiiensis*¹⁰, and the early restriction of germ layer identity (Figure 4.1)¹, I wanted to sequence cells from embryos as early as the S4 stage (8 cells). To understand the gene expression changes happening between the S4 and S6 stage, I hoped to generate a single cell RNA sequencing library of those stages. The manual isolation of single blastomeres of *P. hawaiiensis* was previously achieved by Cassandra Extavour⁴¹. Therefore, I decided to apply the cell picking and plate-based single-cell RNA sequencing technique, CelSeq2⁴². I dissected and dissociated embryos as young as the S4 stage (8 cells) with the help of Cassandra Extavour to separate blastomeres with an eyelash and up until S6 stage (the aggregation of the germ disk) by placing the dissected embryos in a 500ul Eppendorf tube and flicking it. I picked 6 plates (each plate contained a variable number of embryos, depending on the availability of cells after dissociation) for a total of 576 cells in this fashion and kept them at -80C. I then lysed the cells, purified and reverse transcribed their mRNA to generate cDNA libraries using the Cel-Seq2 protocol⁴². I repeated this procedure eight times, but none of these eight attempts at creating libraries were successful as can be seen in Figure 4.12. After the linear amplification step, products from the reverse transcription were fragmented to a 300-1000 bp size prior to the generation of the cDNA library. In my hands, all fragmentations led to a distribution of cDNA molecules length centered around 80bp (Figure 4.12).

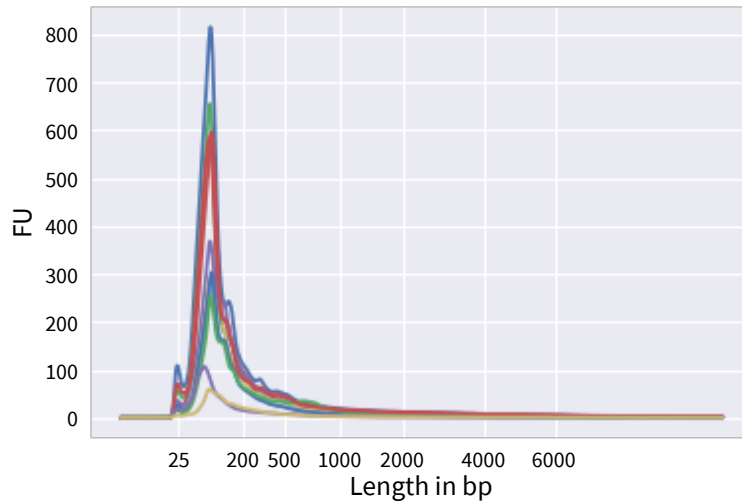


Figure 4.12: Bioanalyzer traces of 8 CelSeq2 library attempts. Each color is one of eight attempts to generate cDNA libraries from isolated blastomeres. The traces were measured of the fragmented amplified RNA after bead cleanup. The main peak in all of them is around 80bp, whereas the expected peak, according to the CelSeq2 protocol⁴² is between 300-1000bp. The Y-axis is Fluorescent Units, a measure of the quantity of DNA flowing through the bioanalyzer elution column at a given size.

4.3.3 Development of an embryo cell dissociation protocol for *P.*

hawaiensis

To sequence embryos older than stage S6, I attempted to apply the inDrop microfluidic encapsulation device⁴³. inDrop requires a highly concentrated (>10000 cells/ml) solution of well-dissociated (>95% single cells), highly viable cells (>95% survival rate) (Allon Klein, Harvard Medical School Single Cell Core (<https://singlecellcore.hms.harvard.edu/>), personal communication). The manual separation of cells used above was not a viable technique to generate such a cell suspension, because cells beyond stage S6 formed a cohesive structure that would not dissociate with the tube flicking method employed for CelSeq2. Therefore I aimed to develop a new protocol for cell dissociation that would be compatible with inDrop⁴³.

I assumed that *P. hawaiensis* blastomeres would contain adherens junctions, a crucial component of which are Cadherins⁶¹. Cadherin-mediated cell adhesion is mediated by the presence of Ca²⁺ ions. I also considered previous reports that enzymatic cell dissociation at high temperatures can stress cells and change their transcriptomic state⁶². Therefore, I aimed to develop a protocol that would maintain the cells at 4°C and avoid the use of proteases. I started from a protocol kindly provided by James A. Briggs (Harvard Medical School) that was

developed for the dissociation of *Xenopus laevis* embryos²⁷. This protocol involved treatment with pronase to remove the vitelline membrane of *X. laevis* eggs²⁷. Upon pronase treatment, the *P. hawaiiensis* eggshell remained intact and showed no sign of softening or digestion with up to two hours of treatment at 37°C (data not shown). To my knowledge, the only tested chemical known to digest the eggshell of *P. hawaiiensis* embryo is bleach (personal communication with Anastasios Pavlopoulos). However, I feared that upon contact with a solution of bleach, cells in the embryo would suffer significant damage. To my knowledge, no studies have developed a successful enzymatic method for crustacean eggshell digestion⁶³. I therefore manually removed the eggshell using tungsten needles. This step limited my capacity to work sufficiently rapidly to dissociate large numbers of embryos. Moreover, the manual dissection of the eggshell results in some damage to the embryo, resulting in the loss of some cells. However, this was the most successful methodology I could find to dissociate a large number of embryos into single cells. With practice, I was able to dissociate 160 embryos in 90 minutes (1.7 embryos per minute).

Once the embryos were removed from their eggshells, I took extra care to remove any extra debris, as I noted that the dissociated cells tended to aggregate around debris particles and did not dissociate properly (data not shown). Similarly, Briggs et al.²⁷ noted that failure to remove all eggshell fragments in the dissociation of *X. laevis* embryos resulted in an aggregation of cells.

Upon transferring the embryos, free of their eggshells, into the dissociation buffer, I noted that the embryos dissociated into individual cells (data not shown). This may have been caused by the sequestration of Ca²⁺ by CAPS (3-(Cyclohexylamino)-1-propanesulfonic acid) in the medium. However, without strong mechanical agitation, I noted that the cells did not fully dissociate, and that doublet and triplet cell aggregates remained (7.7% of doublets and 2.4% of triplets, n=3 dissociation experiments performed for each with 30 S11 staged embryos).

Therefore, I subjected the cells to mechanical agitation as follows: I immersed cells in 7 ml of dissociation buffer in one well of a 12 well plate. The well was sealed such that no air was trapped within it. This prevented cells from coming into contact with air bubbles, which was important to avoid because I noted that, if air bubbles were trapped in the column, almost no cells could be detected in a well after agitation (data not shown). I believe that this is because the surface tension between the cell and a water-air interface leads to the destruction of these

blastomeres, by tearing of the cell membrane. After sealing the wells, I vortexed the plate for 20 minutes. This resulted in a very dilute cell suspension (<1000 cells/ml, the technical lower limit of the hemocytometer used), but the single cell dissociation efficiency increased, when compared to no mechanical dissociation, to 95.4% (with 4.3% of doublets and 0.3% of triplets, n=3), and the viability of the cells was high (96.5%, n=3 dissociation experiments performed for each with 30 S11 staged embryos).

The low concentration of cells (<1000/ml) that I obtained with my method could not be used for inDrop. Therefore, I tried to adapt a methodology to reconcentrate the cell suspension based on the protocol of Briggs and colleagues²⁷. After dissociation, I let the plate rest for five minutes to allow the cells to settle at the bottom of the well. Subsequently, I backfilled (150ul of each Optiprep concentration was pipetted at the bottom of the well one after the other) a density gradient column of Optiprep from the bottom to the top, starting from a 5% solution and finishing with a 40% solution, in 5% increments, creating multiple interfaces between the different densities. Cells were lifted from the bottom of the plate when a denser layer was backfilled. Under a stereomicroscope, the cells visible at the interface between two layers could then be aspirated into a syringe for loading in the inDrop device. This method led to a concentration of 5000 cells/ml (n=3 dissociation experiments performed for each with 30 S11 staged embryos).

Using this technique, I performed a first experiment (referred to as First experiment) at the Harvard Medical School Single Cell Core (<https://singlecellcore.hms.harvard.edu/>). Embryos were collected and staged such that stage S6-7 (12-24hpf) and stage S11 (64-68hpf) embryos were dissociated for an inDrop experiment. The embryos were split into two replicates. Following dissociation, 200 ul of cell suspension, containing approximately 1000 cells (pooled from 30 embryos), were encapsulated for library preparation per replicate (n=2 per time point) (The core staff counted the cells flowing in the inDrop device. However, I believe, after discussing with them, that they counted yolk granules as well inflating the number). Although this implies that my cell suspension was at a concentration of 5000 cells/ml, the results of the subsequent sequencing (discussed below) suggested that the cell concentration was in fact much lower, as fewer than 1000 cells were detected for each of the four libraries prepared in this experiment

(Table 4.1). The protocol was then further refined to increase the cell concentration, as the first attempt was under the lower end of the technical feasibility of the inDrop microfluidic device.

To increase the total number of cells, I changed the concentration methodology as follows: First, instead of letting the cells settle at the bottom of the well following the mechanical separation step, I transferred the liquid column containing the dissociated cells into a 15ml falcon tube. This approach removed the limitation of using only one well to dissociate embryos, as multiple wells (less than 15 ml of total solution) could be pooled into the 15 ml falcon tube. After collecting the suspension, I created a small gradient density by backfilling different concentrations of Optiprep at the bottom of the falcon tube (see detailed protocol below). Finally, the tube was spun in a centrifuge to concentrate the cells, which were then resuspended in 100ul.

A second inDrop experiment (referred to as Second experiment) was then performed at the Harvard Medical School Single Cell Core (<https://singlecellcore.hms.harvard.edu/>) using this new concentration protocol. For stage S11 (64-68hpf) I dissociated 160 embryos. The embryos were again split into two technical replicates of 80 embryos each. Fewer than 160 embryos could be collected for the S6-7 time point, and only one replicate of 80 embryos was done. Following the dissociation, I encapsulated approximately 600 cells for the single S6-7 replicate, 1000 cells for the first S11 replicate, and 2000 cells for the second S11 replicate. I estimated the number of cells encapsulated by making and observing a time-lapse video of the cell suspension passing through the inDrop device under the compound microscope (data not shown).

In total, two inDrop experiments were performed, with four conditions for the first one and three for the second one. The libraries from the first experiment are marked as I SX.replicate and from the second one as II SX.replicate. The resulting libraries were sequenced by the Harvard Bauer sequencing core on the NextSeq platform. The detailed final protocol for the dissociation of *P. hawaiiensis* embryos can be found at the end of this chapter.

Library name	Developmental Stage	# of embryos	Estimated number of encapsulated cells	Total number of reads	Reads after QC	Proportion of reads	Expected proportion
First experiment							
I S6_7.1	S6-7 12-24hpf	30	1000	107102966	88895461	26.29%	25%
I S6_7.2	S6-7 12-24hpf	30	1000	81350942	69148304	19.97%	25%
I S11.1	S11 64-68hpf	30	1000	80390297	67447459	19.73%	25%
I S11.2	S11 64-68hpf	30	1000	112067856	99740391	27.51%	25%
Second experiment							
II S6_7.1	S6-7 12-24hpf	80	600	67693910	54493597	18.12%	25%
II S11.1	S11 64-68hpf	80	1000	84664503	57233204	22.66%	25%
II S11.2	S11 64-68hpf	80	2000	204487930	166453175	54.72%	50%

Table 4.1: List of all inDrop experiments executed on dissociated *P. hawaiiensis* embryos. Each library is displayed as a row with its sample name. The number of encapsulated cells along with the library's summary statistics are shown in the following columns. The estimated number of cells encapsulated was calculated by watching cells flow through in the inDrop device over a set time window of one minute, and extrapolated for the entire experiment. For the first experiment, the measure was done by staff members of the Harvard Medical School Single Cell Core (<https://singlecellcore.hms.harvard.edu/>). For the second experiment, I performed the counting. The Proportion of reads is the number of reads with the index from one inDrop run divided by the total number of reads in that library. The expected proportion corresponds to the mixing of each single library on the same NextSeq lane that was done prior to the sequencing.

4.3.4 Pilot analysis of the transcriptomic profile of cells from S6-7 and S11 stage embryos

To assess the sequencing quality of each library, I ran quality control analyses using FastQC⁶⁴ on the two libraries generated above. The proportion of reads coming from each inDrop run deviated only slightly from the expected count (expected proportion of reads derived from each library based on the library mixing performed prior to sequencing) (Table 4.1, Proportion of reads vs Expected proportion).

FastQC reported an over-representation of the specific sequence

ACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTTC. This represented 14.12% of the reads in the first experiment and 11.64% of reads in the second

experiment. A BLAST⁶⁵ search for this sequence on the genome Phaw_5.0 (NCBI ID: GCA_001587735.2) with default parameters yielded no hits, but an NCBI BLAST search⁶⁶ against the non-redundant (nr) database with default parameters recovered two hits, the first against the *P. hawaiiensis* complete mitochondrial genome (NCBI ID: NC_039402.1) and the second against the *P. hawaiiensis* partial mitochondrial genome (NCBI ID: AY639937.1). After running NCBI blasts on nr for the 63 overrepresented sequences (listed at the end of the chapter in: *List of overrepresented sequences in the two libraries*) detected by FastQC, all the reads mapped to the *P. hawaiiensis* mitochondrial genome. I therefore concluded that these reads were likely of mitochondrial origin, and removed them by mapping all reads against the previously published *P. hawaiiensis* complete mitochondrial genome (NCBI ID: NC_039402.1) and retaining only the unmapped reads. The full results of the filtering are displayed in Table 4.2.

I believe that the large amount of mitochondrial RNA (mtRNA) may be due to cell damage during the dissociation protocol, causing mitochondria to enter the suspension along with the cells. The mitochondria could in principle have been encapsulated into droplets, in which case some droplets should contain only mitochondrial DNA, and others might contain a mixture of both mitochondrial and genomic DNA (as cells contain mitochondria and a proportion of ~20% mtRNA was reported previously^{27,29}). An alternative hypothesis is that embryonic cells contain a large amount of mtRNA, in which case every droplet would display a high amount of mtRNA reads. Plotting the per droplet proportion of reads matching mitochondrial genes versus the proportion of reads matching nuclear genes did not distinguish clearly between these hypotheses (Figure 4.13). For the first experiment, I detected a continuum between droplets containing >90% mtRNA to droplets containing >90% genomic mRNA, with 57% of droplets containing >70% mtRNA. For the second experiment, each of the three runs displayed a different profile. II S6_7.1 showed a similar distribution as I S11.2 (mtRNA proportion for II S6_7.1 average: 45%, standard deviation: 22%; I S11.2 average 43%, standard deviation: 22%). II S11.1 had 69% of the droplets with more than 60% of the reads mapping to the nuclear genome, a result that is also evident in Table 4.2, which shows that 33% of the reads from II S11.1 map onto the mitochondrial genome. Finally, II S11.2 showed a distribution with only 3% of the droplets with more than 60% of the reads mapping to the nuclear genome. However, this low

Library Name	Developmental Stage	# Reads after QC	Non-mtRNA	% mtRNA	% Genome	# droplet barcodes	# droplet after QC "cells"
First experiment							
I S6_7.1	12-24hpf	89M	19M	78%	18%	2709	8
I S6_7.2	12-24hpf	69M	8M	88%	10%	446	15
I S11.1	60-64hpf	67M	18M	81%	16%	1726	60
I S11.2	60-64hpf	100M	23M	77%	20%	650	112
Second experiment							
II S6_7.1	12-24hpf	54M	16M	70%	25.90%	1091	405
II S11.1	60-64hpf	57M	38M	33%	58.90%	2359	1344
II S11.2	60-64hpf	166M	69M	58%	34.50%	68784	607

Table 4.2: Summary statistics of the processing for each inDrop experiment. The sample and library names are provided in the first two columns. # Reads after QC present the statistics for the number of reads after quality control. Non-mtRNA is the number of reads mapped on the nuclear genome. % mtRNA is the proportion of reads mapping to the mitochondrial genome. % Genome is the proportion of reads mapping to the nuclear genome. # droplet barcodes is the number of detected droplet barcodes in a library. Droplets containing a Cell is the total number of cells kept after droplet quality control.

proportion is due to the much larger number of droplet barcodes, 68784 versus 2359 for II S11.1 and 1091 for II S6_7.1 (Number of barcodes for all libraries in Table 4.2). The very high proportion of reads mapping to the mitochondrial genome implies that the first experiment resulted in mostly mitochondria being encapsulated, proportionally reducing the signal coming from encapsulated cells to the point that fewer than 23M reads per experiment remained (Table 4.2). The differences between experiments in the cell concentration technique, as discussed above, could explain this discrepancy.

To try to retain only the droplets that likely contained a cell (and not just mitochondria) for analysis, I removed the reads deriving from any droplet that had more than 20% of its reads mapping to the mitochondrial genome and fewer than 50% of its reads mapping to the nuclear genome. This process is referred to as droplet Quality Control or droplet QC. The droplets meeting these criteria were called "cells", and their numbers for each library are displayed in Table 4.2. The first experiment yielded a total of 195 "cells" across all four runs (compared to a predicted 4000 cells). The second experiment yielded 2361 total cells across all three runs (compared to a predicted 3600 cells). Therefore for the subsequent analysis, I did not analyze data from the first experiment, as almost no usable data remained following my attempt to correct for likely mitochondrial contamination.

To analyze the gene expression pattern of the remaining droplets, I needed to map each read to

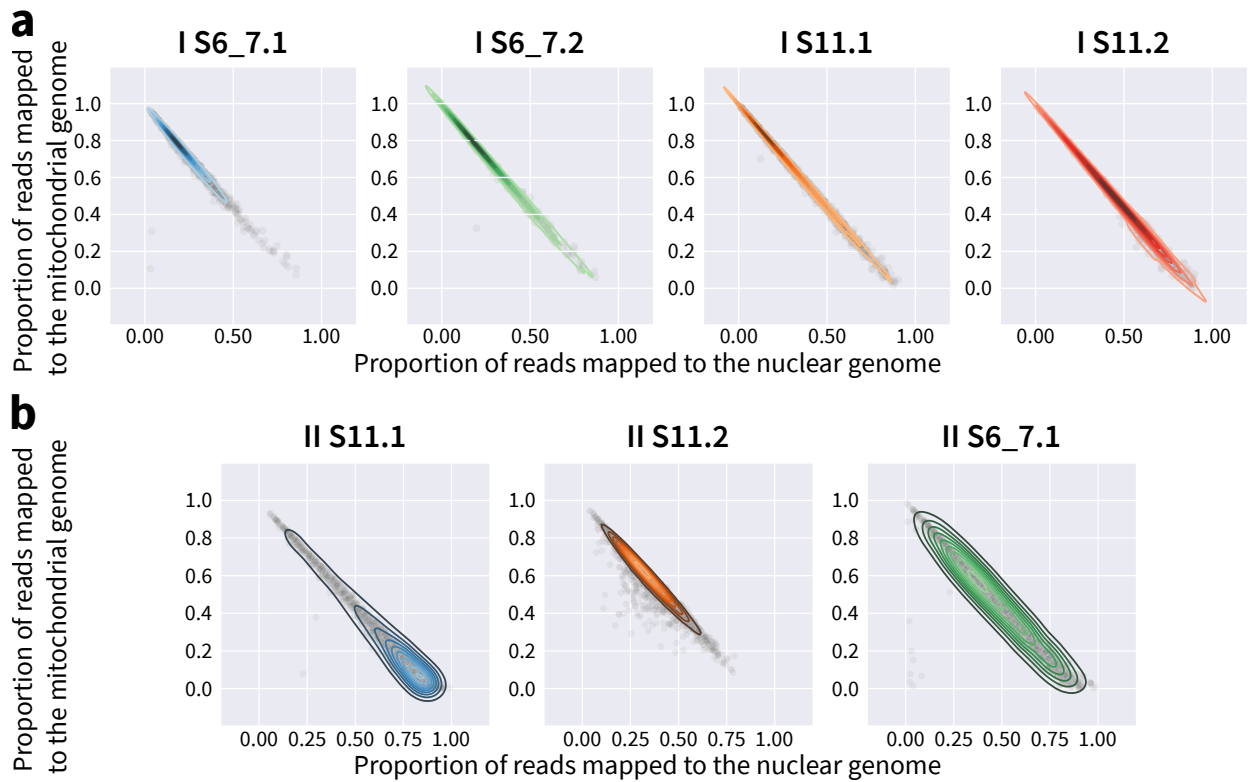


Figure 4.13: Proportion of reads per droplet that mapped to the mitochondrial genome and the nuclear genome of *P. hawaiiensis*. Each droplet corresponds to a unique cell barcode sequenced. The distribution of droplets was fitted using a Kernel Density Estimation (KDE) and the distribution was drawn above the droplets using a contour plot. **a)** Reads from the first sequencing library coming from the first four inDrop experiments. **b)** Reads from the second sequencing library coming from the second three inDrop experiments.

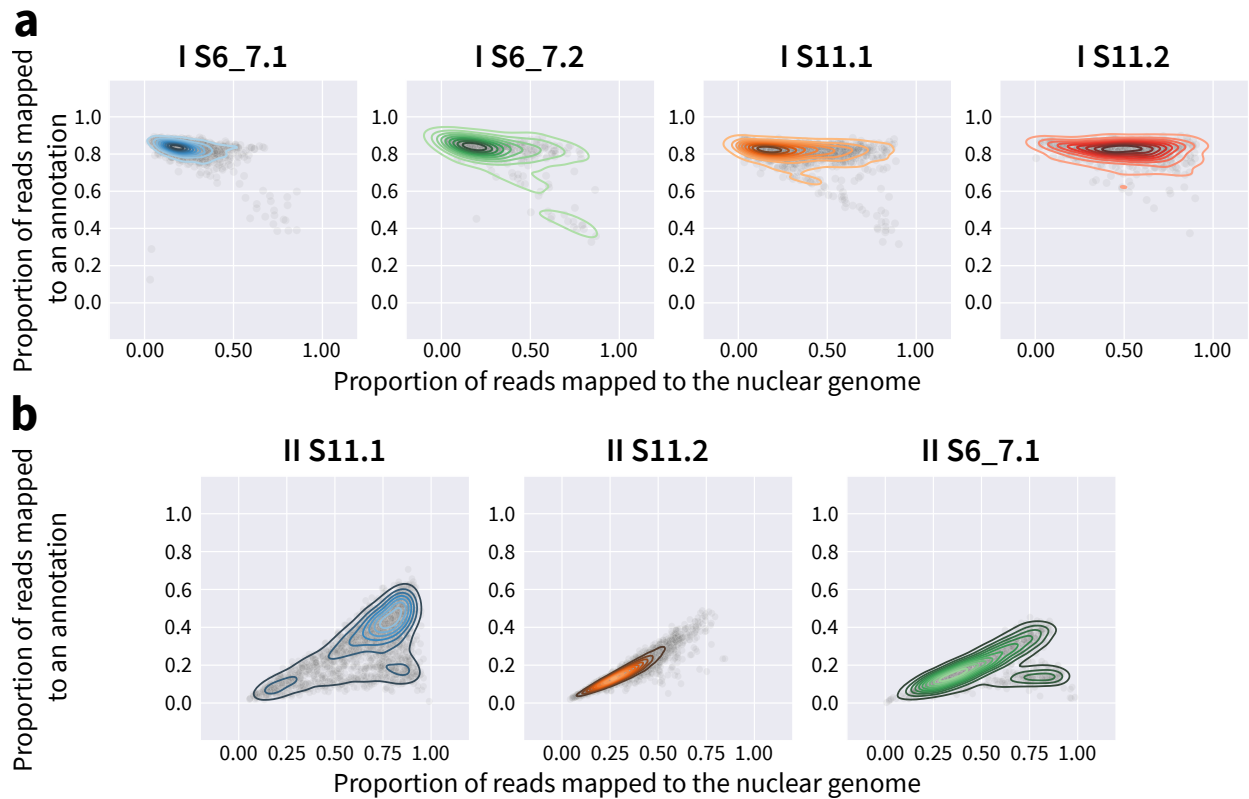


Figure 4.14: Proportion of reads per droplet that mapped onto a gene annotation against the reads that mapped to the nuclear genome of *P. hawaiiensis*. Each droplet corresponds to a unique cell barcode sequenced. The distribution of droplets was fitted by a Kernel Density Estimation (KDE) and the distribution was drawn above the droplets using a contour plot. **a)** Reads from the first sequencing library coming from the first four inDrop experiments. **b)** Reads from the second sequencing library coming from the second three inDrop experiments.

a predicted transcript or gene. The *P. hawaiiensis* nuclear genome annotation set used for this analysis was kindly provided by Damian Kao (University of Oxford, UK). This annotation was based on adult RNA transcriptome data, and therefore lacked de novo or non-coding annotations. The proportion of reads that mapped to an annotation varied across libraries. Droplets in II S6_7.1 displayed a bimodal distribution, with 56% of droplets having ~35% of the reads mapping to an annotation, and 44% of droplets having ~15% of the reads mapping to an annotation (Figure 4.14). For II S11.1 and II S11.2, 59% of droplets had between 40% and 50% of their reads mapping to an annotation, however, 12% of droplets displayed ~15% of their reads mapping to an annotation (Figure 4.14).

The inDrop protocol performs the reverse transcription step with a primer tagged with a Unique Molecular Identifier (UMI)⁴³. In the "cells" remaining after the filtering steps described above, I computed the number of unique UMI per cell, and used this number as a proxy for the depth

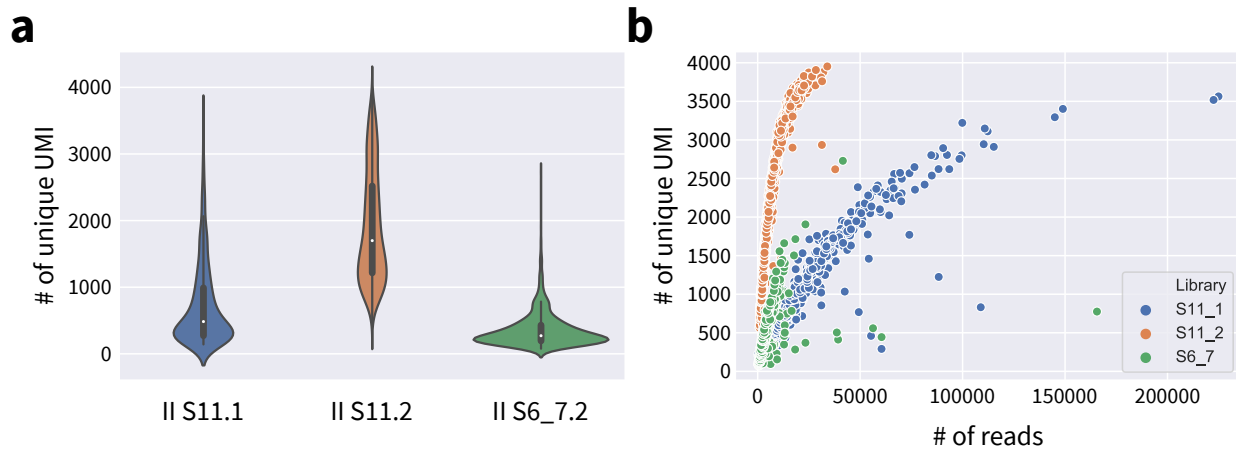


Figure 4.15: Analysis of the number of unique mRNA molecules captured by inDrop. a) Violin plots of the number of UMI in each droplet from the three inDrop runs performed during the second experiment. **b)** Saturation plot for the number of UMI per cells. The number of UMI in each cell is plotted against the number of reads sequenced for that droplet.

with which a particular cell was sequenced. To determine the expected number of UMI per cell, I performed a saturation analysis by plotting the number of unique UMIs against the number of reads. In other reported droplet-based single cell RNA sequencing experiments^{27,29,43}, the UMI saturation curve converges asymptotically to a value that is the maximum expected amount of UMI's in a cell. For the second experiment, none of the runs displayed a complete curve, so that I was unable to accurately measure the saturation point (Figure 4.15). However, II S11.2 and II S11.1 showed inflections that might suggest a saturation point between 4000 and 5000 unique UMIs (Figure 4.15b). By plotting the distribution of unique UMI counts per run, I found that the II S11.2 run showed the highest number of unique UMIs (median 1699), whereas II S11.1 had a 484 median UMIs, and II S6_7.1 had a median of 271 UMIs (Figure 4.15). Thus, the sequencing depth of each cell was lower than the expected 4000-5000 UMIs per cell. Interestingly, the depth at which cells from II S11.2 were sequenced was much higher than those from II S11.1, despite the mtRNA content being much higher in II S11.2.

Based on this analysis, I removed data from droplets with fewer than 100 UMI. After this filtering, 1337 cells remained in II S11.1, 537 in II S11.2, and 402 in II S6_7.1.

Using these admittedly sparse data, I ran pilot analyses aimed at testing multiple hypotheses:

- Hypothesis 1: Four main cell clusters would be detectable at stages of S6-7, each one corresponding to a germ layer.

- Hypothesis 2: The diversity of cell types would be larger at stage S11 than at Stage S6-6.
- Hypothesis 3: Some cell populations would be shared by both time points.
- Hypothesis 4: Putative cell types could be identified, and gene markers could be used to assess their germ layer identity.

RNAseq generates datasets that contain thousands of cells and tens of thousands of genes⁴³. This creates high dimensionality matrices which in turn can become problematic due to, for example, the distance metric tending towards 0 when the number of dimensions increases⁶⁷. Therefore, one of the common ways to preprocess such a dataset is to reduce its dimensionality by filtering out dimensions contributing little information⁶⁷. As such, the 21763 genes (65% of the total of 33385 detected genes) that were expressed by fewer than 3 cells were removed from the matrix. Finally, the counts were normalized as previously described in Weinreb et al.⁶⁸.

I computed cell clusters on the II S6_7.1 dataset using the Leiden algorithm⁶⁹. I then computed a UMAP embedding on the cells⁷⁰, and plotted each cell on the resulting two-dimensional space (Figure 4.16). Three main clusters of cells appeared (Figure 4.16). One of those clusters, Cluster 2, was composed of cells that had a smaller distance between them than the distances between the cells in the other two clusters (Figure 4.16). This suggests that the cells within Cluster 2 have similar transcriptional states.

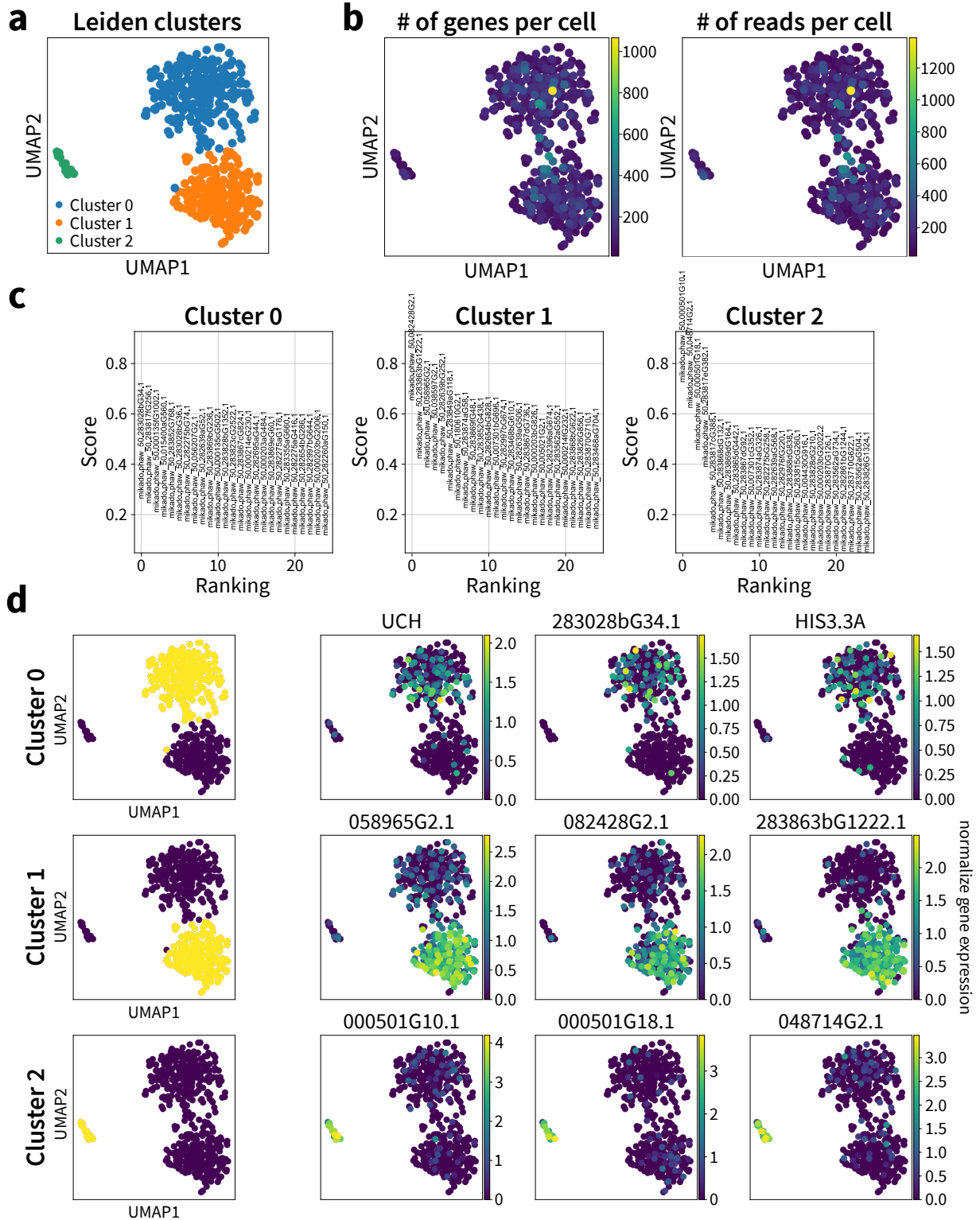
To determine whether the apparent cluster separation was due solely to the number of genes or number of reads per cell, I colored each cell on the UMAP embedding by those two metrics. In both cases, visual inspection did not suggest that any clusters displayed enrichment in any of those values (Figure 4.16). I then asked whether clear sets of gene markers for each cluster could be inferred by fitting a logistic regression model for each cluster⁷¹. Cluster 0 did not show a clear set of genes. Cluster 1 and 2 however showed a set of genes that were only expressed in those cells (Figure 4.16). This result can be visualized by coloring the cells on the UMAP embedding according to the normalized expression value of each gene. A similar result was obtained by looking at the top three marker genes of the three clusters: cells from Clusters 1 and 2 had a strong differential signal compared to Cluster 0 (Figure 4.16). None of the gene markers

for those clusters matched any known protein. Thus, I cannot speculate as to the potential germ layer or other differentiated identities of those clusters based on these data alone. I cannot definitively conclude that at the S6-7 stage, these scSeq data can differentiate between three cell populations.

I repeated the analysis above in the exact same fashion for the II S11.1 and II S11.2 samples, merging both samples into a single dataset. The Leiden clustering algorithm predicted 15 clusters⁶⁹ (Figure 4.17). Visual inspection of the distribution of gene and read counts on the UMAP embedding did not suggest any clear pattern; the read counts appeared evenly spread across the different clusters (Figure 4.17). The gene marker detection analysis⁷¹ did not show a strong step decrease in the score for most clusters (Figure 4.17). To better understand how the gene expression varied between clusters, I plotted heatmap and dendrogram representations of the clusters (Figure 4.18). Consistent with the results of the gene marker analysis⁷¹, there did not appear to be a clear set of genes differentially expressed between the clusters, with the exception of clusters 2, 7, and 12, which appeared to have a stronger unique expression of genes than other clusters (Figure 4.18). To conclude, the analysis of the S11 datasets did not lead to clear segregation of the cells into subpopulations.

Figure 4.16 (following page): Preliminary analysis of the pilot scRNAseq data for II S6_7.1. Each cell is represented in a 2D UMAP embedding. **a)** Results of the Leiden clustering algorithm (a clustering algorithm tailored to scRNAseq data⁶⁹) are shown as different colors per cluster. Three clusters were found. **b)** Visualisation of the number of unique genes and the number of reads for each cell on top of the UMAP embedding. Each cell was colored using a gradient representing the number of unique genes or the number of reads. **c)** Ranked scatter plot of the top 25 gene markers for each cluster. Genes are ranked in decreasing order of their Logistic regression scores⁶⁸ (The one-sided Z score of the p-value given by the logistic regression model for a given marker restriction within a cluster) (Y-axis). **d)** Visualisation of the top three marker genes (from the analysis shown in c) expression patterns on the UMAP embedding. Their normalized expression level is represented by coloring each cell. The gene name, if the reads mapped to an annotated gene with a putative homology identity, is displayed on top of each plot (applies only to UCH (Ubiquitin carboxyl-terminal hydrolase) and HIS3.3A (Histone H3.3A)). All other marker genes lacked predicted homologs; their unique identifiers from the published genome annotation⁴ are used.

Figure 4.16: (continued)



To compare the gene expression profile between cells from S6-7 embryos and S11 embryos, I merged the II S6_7.1, II S11.1, and II S11.2 datasets and performed a UMAP embedding onto the resulting dataset. First, it appeared that some of the cells of S6_7 embryos separated from the main S11 cell cluster (Figure 4.19). A large cluster of cells from S6_7 and S11 clustered together, away from all other clusters. Finally, a smaller proportion of cells from both S6_7 and S11 embryos clustered tightly together (Figure 4.19). To distinguish differential expression patterns between the S6_7 and S11 embryos I plotted the top 10 genes of each cluster against the different runs. The resulting dendrogram confirmed that the S6_7 library clustered away from the two S11 libraries. Moreover, some gene markers could be distinguished that were only expressed in either S6_7 or S11 (Figure 4.19). Because I hypothesized that the cells from S6_7 embryos would not cluster with S11 embryos due to the high amount of development time between the two time points, I researched the gene markers unique each of the two co-clusters. The larger cluster's most differentially expressed gene was phaw_50.282654bG828.1, an unknown Ubiquitin protease of the USP family. Some other marker genes of this cluster included ribosomal proteins and ubiquitin hydrolases. One hypothesis to explain this observation might be that this cluster is mostly formed by cells with a high translation rate. The second, smaller cluster is composed of unknown proteins with no match in the UniprotKB database⁷². Moreover, this is the same cluster that in the analysis of the S6_7 cells was found to have less distance in gene expression between cells. Once more, the low amount of information in each cell with only a few hundred genes per cell precludes a more thorough analysis of the dataset. Nevertheless, I consider these results encouraging, as even with such a low-quality dataset, the gene expression landscape seems to have changed enough for cells from the earlier embryos (S6_7) to cluster independently of the cells from later embryos (S11).

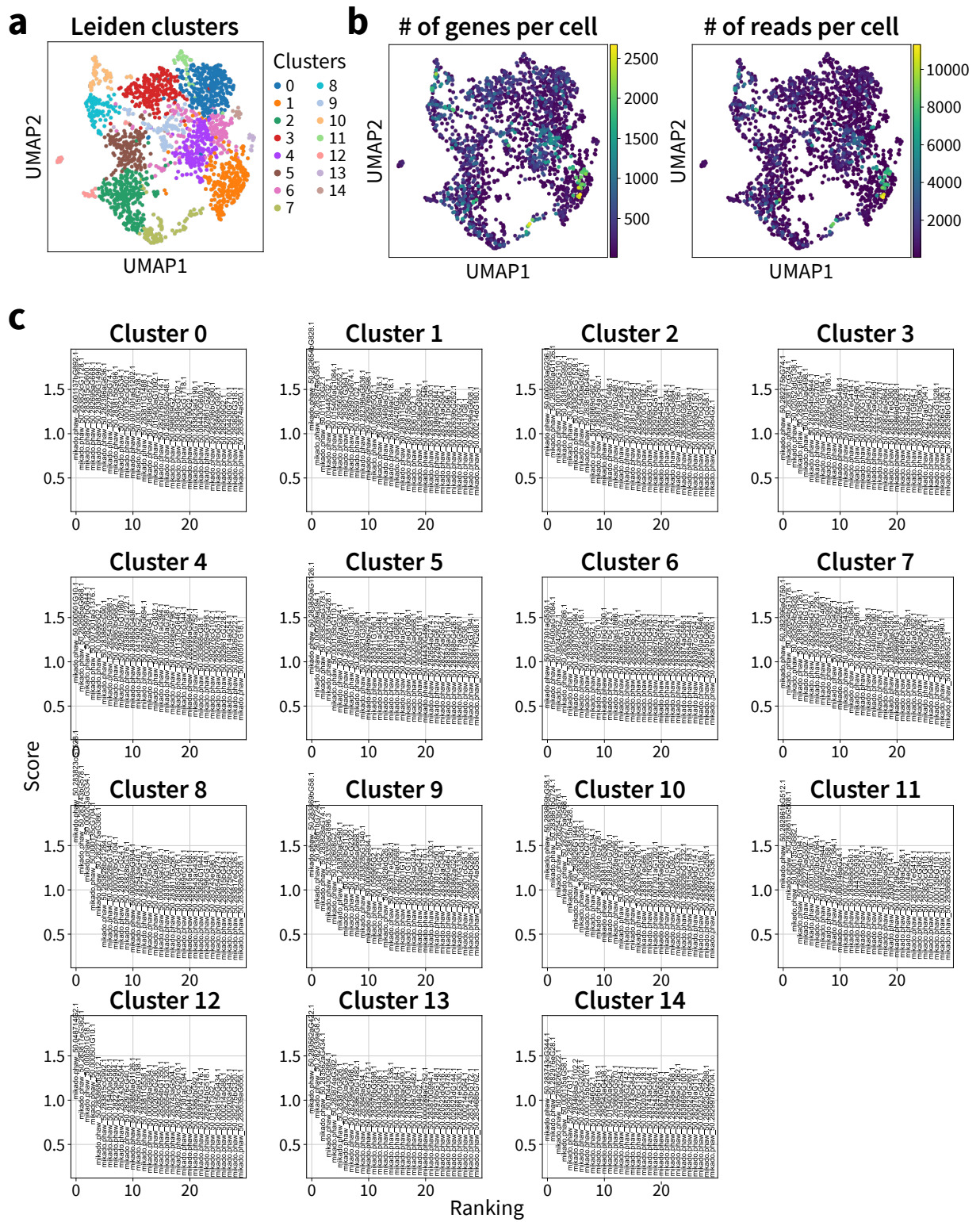
Finally, I wanted to know if I could use known markers of germ layers to identify cells and observe whether clusters were expressing them even at low levels. As a pilot experiment, I collected a set of 84 ectodermal markers expressed during arthropod development and found in the genome of *P. hawaiiensis* (based on publications and GO terms)^{16,73,74}. The ectoderm composes most of the cells of the early embryo of *P. hawaiiensis*, therefore I expected to find a large number of cells expressing ectodermal markers. Out of the 84 putative ectodermal

markers, 42 were expressed by at least one cell in at least one of the three datasets. I then visualized the expression of each gene by coloring the cells on the UMAP embedding by their expression levels. From the 42 genes expressed in my datasets, I selected the 15 with the highest number of cells expressing them for further analysis (Figure 4.20). Four genes were expressed in the S6_7 cluster: *single-minded (sim)*, a ventral midline marker in *P. hawaiiensis*⁷⁵, *paired (prd)* an early embryonic segmentation gene⁷⁶, *groucho (gro)*, a Wnt and TGF-beta downstream effector⁷⁷ and *escargot (esg)* a Snail-type transcription factor involved in morphogenesis⁷⁸ (Figure 4.20). This is interesting in light of an earlier report of the expression of *sim* in later stages of *P. hawaiiensis* embryogenesis, starting at S11 and establishing a strong expression by S14⁷⁵. Here, I detected the expression of *sim* in a small population of cells as early as S6_7 embryos. These results are encouraging as some putative ectodermal markers could be found in a subpopulation of the cells in my dataset. However, the absence of expression of ectodermal markers that I would have expected to be expressed at the S11 stage, such as *eve* (even-skipped) or *distal-less (dll)*⁷⁵, is consistent with the overall low quality of the scRNAseq datasets. Reads mapping to one gene are found in at least some cells of all cell clusters, namely *belle (bel)*, a DEAD-box helicase necessary for *Drosophila* embryo survival, and active in the male germline⁷⁹ (Figure 4.20).

To conclude, the low depth at which the cells were sequenced led me to abandon any further analysis of this dataset, as it would not be not useful for the questions of interest to this thesis, namely, the study of germ layer differentiation. While I obtained some encouraging results, the high mtRNA amount present in the libraries prevented the detection of mRNA reads in the cells at sufficient depth. Until the cell dissociation steps can be further refined, the study of single-cell transcriptomics of *P. hawaiiensis* will be difficult.

Figure 4.17 (following page): Preliminary analysis of the pilot scRNAseq data for the II S11.1 and II S11.2. Each cell is represented on a 2D UMAP embedding. **a)** Leiden clustering results are shown as different colors per cluster. 15 clusters were found. **b)** Visualisation of the number of unique genes and numbers of reads for each cell plotted on top of the UMAP embedding. Each cell was colored using a gradient (colored bar at right) representing the number of unique genes (left plot) or the number of reads (right plot). **c)** Ranked scatter plot of the top 25 gene markers for each cluster (each gene receives a score, and are ranked from lowest to highest, the ranked plots here only display a subset of those genes). For all plots, the X-axis is the ranking of a gene and the Y-axis the Logistic regression score (as in Figure 4.16). Genes are ranked in decreasing order for the Logistic regression scores. Each score is computed for the segregation capacity of a gene against all other clusters.

Figure 4.17: (continued)



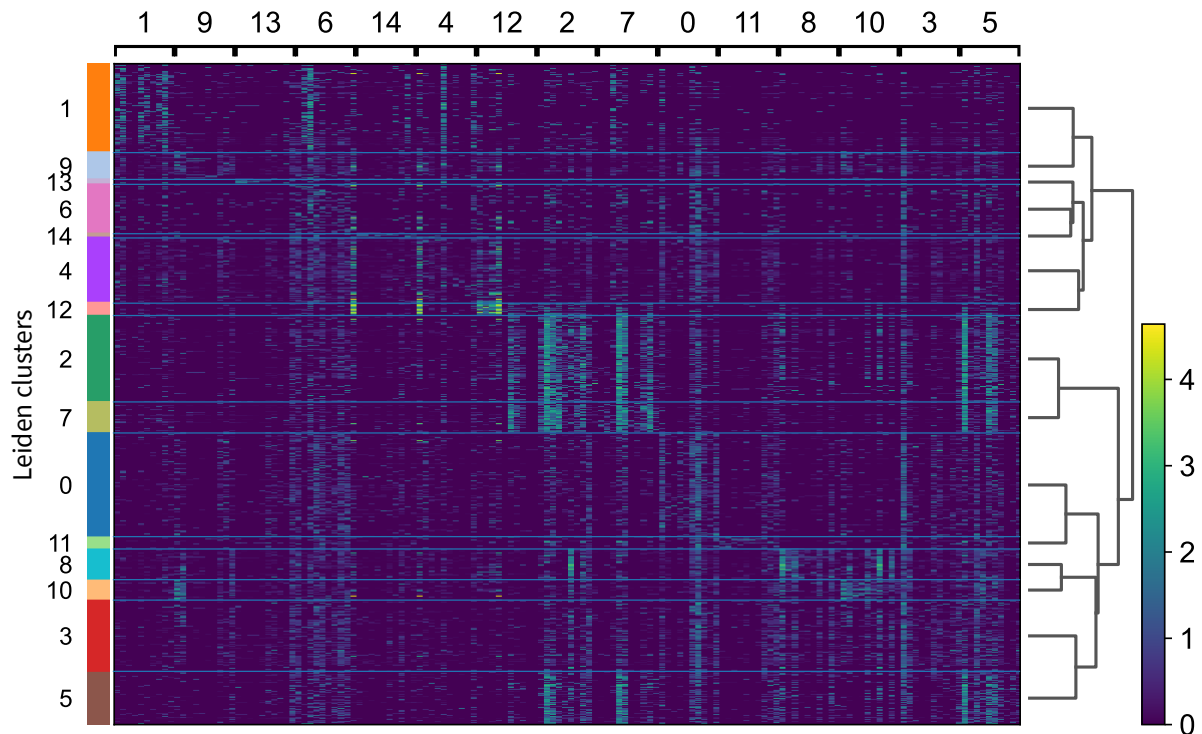


Figure 4.18: Heatmap representation of the expression level for the top 10 marker genes (as per the Logistic regression model performed in Figure 4.17) in each cell cluster predicted from the II S11.1 and II S11.2 datasets. The clusters are displayed based on the computation of the dendrogram displayed on the right and each row represents a cell. Each column represents a gene. Genes are grouped in their clusters based on their ranking in the logistic regression model. Clusters are colored using the same colors as in Figure 4.17. On the right, a dendrogram representation showing the results of a parsimonious tree computation. The distance between each cluster is measured from the gene expression levels and used to compute this tree. The color of the heat map represents the normalized gene expression level and corresponds to the heatmap on the right.

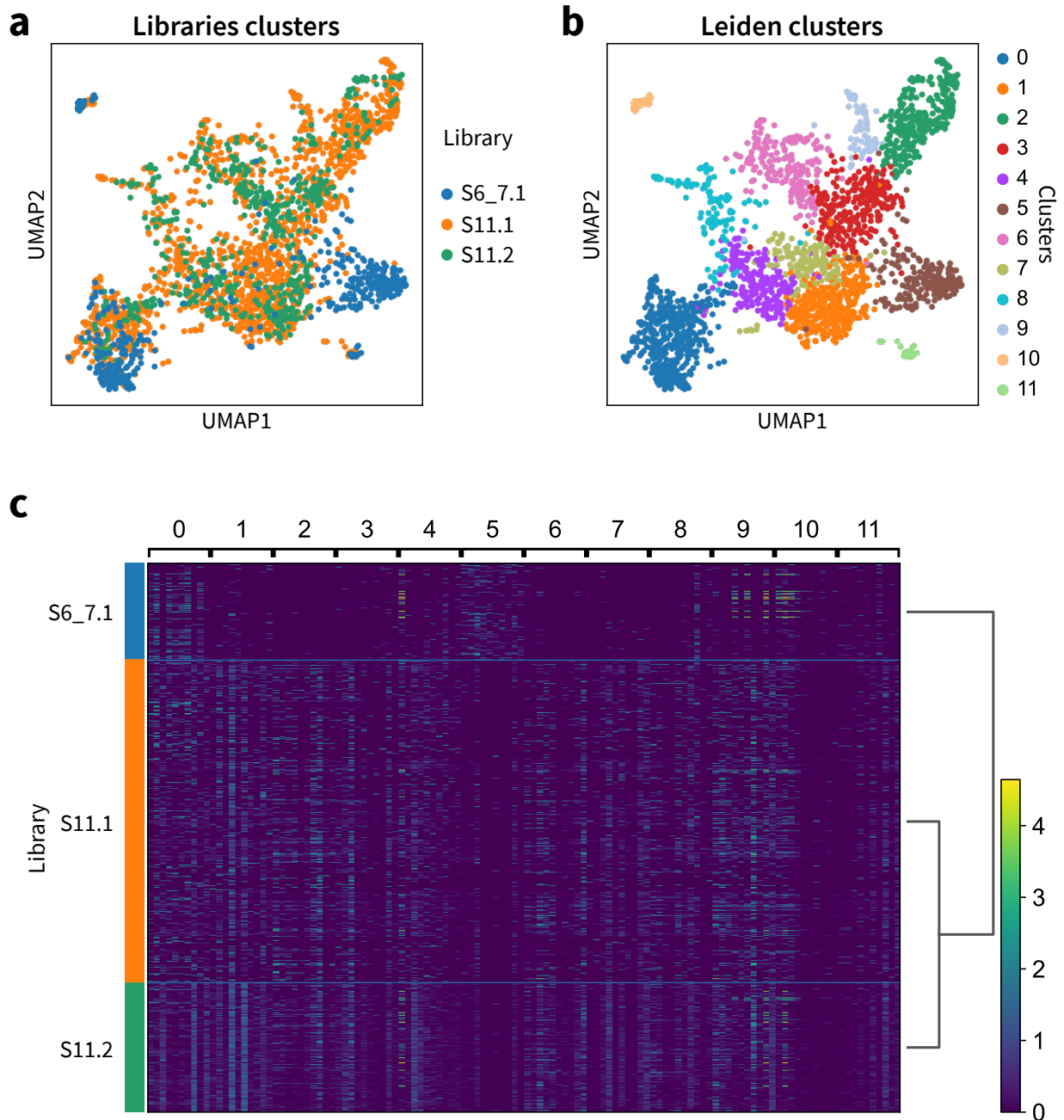


Figure 4.19: Comparative analysis of the scRNAseq data for the cells coming from S6, S7 embryos (12-24hpf), and the later stages S11 (64-68hpf) embryos. Each cell is represented on a 2D UMAP embedding. **a)** UMAP embedding colored by cell library. Both S11 libraries overlap mostly, while S6_7 clusters slightly apart. **b)** Leiden clustering results are shown as different colors per cluster. 12 clusters were found. **c)** Heatmap representation of the expression level for the most discriminant 10 genes in each cell cluster. The cells are ranked from top to bottom by cluster identity. The corresponding clusters are displayed from left to right. Clusters are colored using the same colors as in b. On the right, a dendrogram representation showing the results of a parsimonious tree computation. The distance between each cluster is measured from the gene expression levels and used to compute this tree.

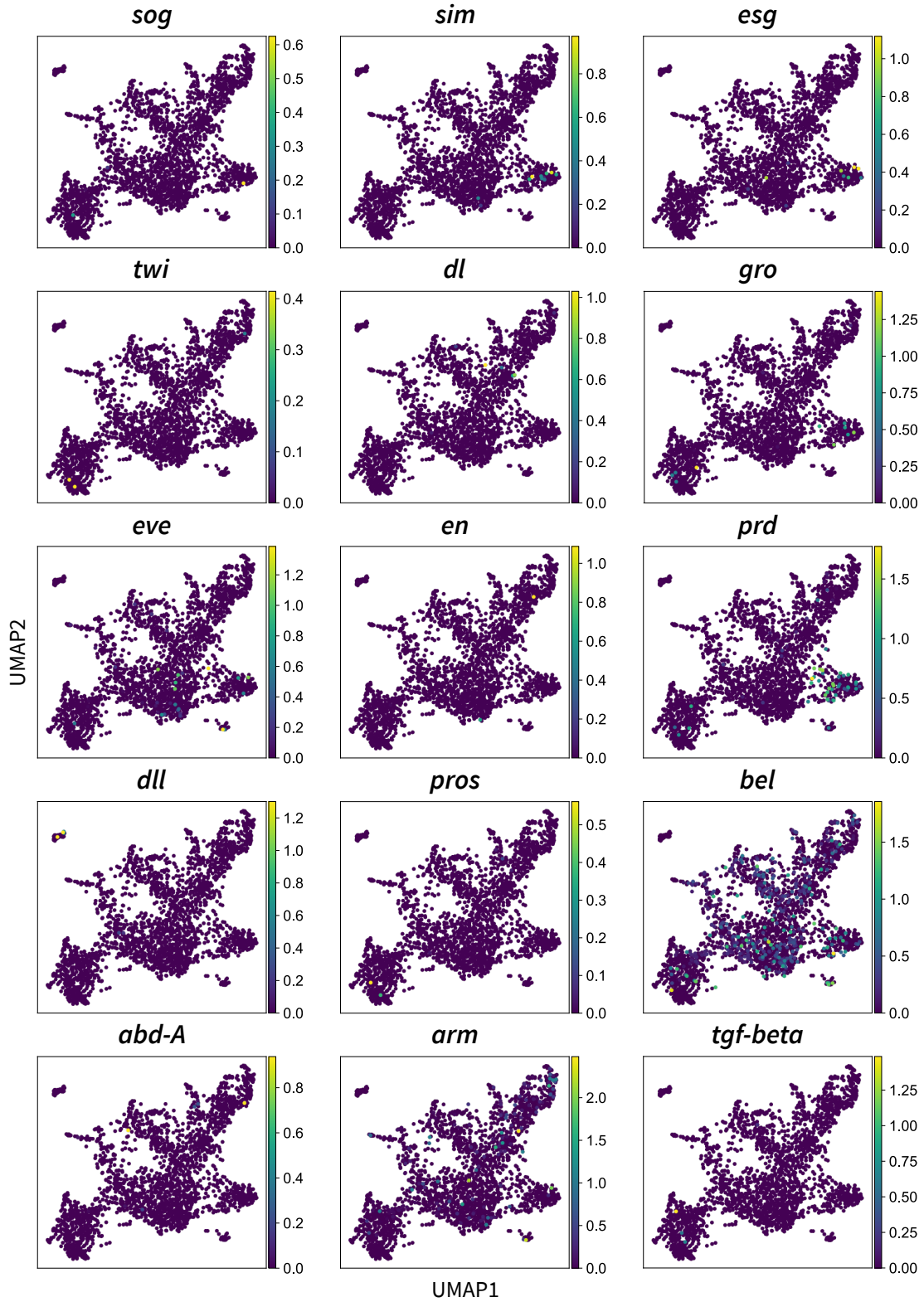


Figure 4.20: Expression level analysis for Ectoderm gene markers collected from the literature. Visualization of the gene's expression patterns on the UMAP embedding computed in Figure 4.19. For each marker, its normalized expression level is represented by coloring each cell. The gene name is written atop each plot using the *D. melanogaster* gene symbol nomenclature.

4.3.5 Annotation of the *P. hawaiiensis* v5.0 genome

The first publication of the *P. hawaiiensis* genome reported a highly fragmented genome assembly⁴ with an N50 of 81,190 bp. Therefore, four laboratories that study *P. hawaiiensis* (Nipam Patel's laboratory, Marine Biological Laboratory, WoodsHole USA; Michaelis Averof's laboratory, Institut de Génomique Fonctionnelle de Lyon, France; Anastasios Pavlopoulos' laboratory Institute of Molecular Biology and Biotechnology, Greece; Cassandra Extavour's laboratory, Harvard USA; unpublished data) collaborated to resequence the genome using the Dovetail platform⁸⁰, resulting in a new version of the genome, Phaw_5.0 (NCBI ID: GCA_001587735.2). This new version displayed an N50 of 20,228,728 bp, almost 250 times longer than the initial assembly. However, the new version was unannotated. Because <50% of the reads from my pilot inDrop libraries mapped to the original *P. hawaiiensis* genome annotation (see above), I decided to generate a new annotation set using the Phaw_5.0 genome.

I first tried to annotate the genome using the previously published set of transcripts annotated from the V3.0 genome. However, this led to a mapping of only 1% of my inDrop reads to an annotated region of the genome, while 96.5% of the reads mapped onto unannotated regions. I therefore attempted to improve the genome annotation by collecting published and unpublished sequencing reads from the *P. hawaiiensis* community (reads obtained from Michaelis Averof (Institut de Génomique Fonctionnelle de Lyon, France), Anastasios Pavlopoulos (Institute of Molecular Biology and Biotechnology, Greece), and Ezio Rosato (University of Leicester, UK)), and using the MAKER2⁴⁴ genome annotation pipeline. To quantify the annotation quality I used the BUSCO score metric⁵⁶ against the Arthropod taxonomic set.

The MAKER2 pipeline⁴⁴ was set to use all gene evidence mapped against the genome, external ESTs from available amphipod transcriptomes, and a de novo gene model generated for Augustus by BUSCO⁵⁶ (see Methods: Annotation of *P. hawaiiensis* genome for the full list of parameters used). After running the MAKER2 pipeline for two subsequent rounds, I generated an annotation set with a BUSCO score of 91.5%, containing 24,422 predicted protein-coding genes, and 109,376 other annotations including repeated elements, transposons, etc. In the future, I recommend that this new annotation be used for the re-analysis of the scRNAseq

libraries generated in this thesis.

4.4 Discussion

In this chapter, I presented the first steps towards a highly ambitious goal, the creation of an Atlas of early embryonic development for *P. hawaiiensis*. However, multiple difficulties hindered the completion of this project. I here discuss some of the reasons behind those failures, and summarize the findings and original datasets that I generated.

In the first part of this chapter, I investigated the Maternal to Zygotic Transition (MZT) of *P. hawaiiensis*. My results suggest that some, but not all, cells may begin zygotic transcription by the 16 cell phase of the S5 stage, which is earlier than the 32 cell phase of the S5 stage previously reported¹⁰. Moreover, I observed a mosaic pattern of phosphorylated RNA Pol II transcription in these embryos, consistent with previous reports¹⁰. This mosaic pattern might reflect random zygotic genome activation occurring between the 16 and 128 cell phases of the S5 stage. Alternatively, it might reflect the activation of different lineages at different times. To try to distinguish between these hypotheses, I began to generate a new dataset of light sheet microscopy images of embryos stained for phosphorylated RNA Pol II, but did not complete analysis of these data. Going forward, I would recommend an algorithm such as RANSAC⁸¹ to align embryos of the same stage and analyze the inter-embryo viability of genome activation. If the analysis of this dataset is completed in the future, I believe that it might be possible to correlate the activation of nuclei between embryos, thus providing data to test the hypothesis that zygotic genome activation is non-random and depends on the lineage.

In the second part of this chapter, I discussed the difficulties and successes towards the sequencing of the transcriptome of single embryonic cells in *P. hawaiiensis*. The dissociation of embryos was much more complex than I expected. Because it was impossible to digest the eggshell without damaging the underlying cells, generating a large number of cells was difficult. Nonetheless, I developed a dissociation protocol that generated a concentration of cells at the low end of the technical capabilities for droplet-based scRNAseq. However, despite the apparent viability of the cells, a large number of mitochondria may have been released into the solution.

This led to libraries saturated by mtRNA, which prevented the sequencing depth of cellular mRNA that would have been necessary to produce any meaningful analysis of germ layer specification or cell type differentiation. In this dataset, by 64-68 hours of development, the embryo still lacks a strong differentiation in gene expression. This seems unlikely given that by that stage the specification of each germ layer is established, and establishment of the major polarization axes and segmentation has started. I believe it is more likely that the low sequencing depth for each cell led to a relatively low amount of information captured. Without more information on gene expression, the statistical analysis of this dataset is not possible beyond this point.

I would like to come back to this point and develop a bit further the main issues that to this part of the project not functioning as expected. First, the low number of cells in *P. hawaiiensis* embryos that facilitates accurate tracking, and, in theory, would make the problem of mapping single-cell RNA sequencing results to light sheet data a less complex problem, is a disadvantage for single cell RNA sequencing given that dissociation of embryos requires manual dissection of the eggshell. I was able to dissect 1.7 embryos per minute, which was not enough to produce the number of cells needed. Second, the fact that the highest quality available version of the genome was not well annotated, required an extensive amount of work to generate new annotations. I cannot stress enough the fact that this endeavor would not have been possible without the unfettered support from the members of the Michaelis Averof (Institut de Génomique Fonctionnelle de Lyon, France), Anastasios Pavlopoulos (Institute of Molecular Biology and Biotechnology, Greece), and Ezio Rosato (University of Leicester, UK) laboratories, who shared unpublished reads to allow the new annotation to be as accurate as possible. I have shared this new annotation with the *Parhyale* community and I hope that it will be helpful to other researchers working on similar projects.

Despite the multiple problems encountered, I believe that it would still be possible to use a version of the protocol I developed here towards the creation of a *P. hawaiiensis* Atlas. Indeed, thanks to CRISPR-Cas9 technology, it is now possible and inexpensive to remove sequences from a cDNA library^{82,83}. By creating a set of guide RNAs targeting the most common *P. hawaiiensis* mtRNA sequences, it might be possible to reduce their concentration by breaking them down

prior to the final amplification step^{82,83}. Using this method could potentially artificially increase the depth of sequencing for the remaining cells. However, due to the overall low encapsulation of cells, multiple subsequent rounds of dissociation and encapsulations would be required, increasing the overall cost of the experiment.

To conclude, I developed part of the foundations required for the generation of an Atlas for *P. hawaiiensis*, which can serve as a starting point for future work in this area.

4.5 Detailed protocol for *P. hawaiiensis* embryo single-cell dissociation

For the whole protocol, work at 4°C (this includes any dissection and dissociation steps that must be done in a cold room), and place all solutions on ice.

4.5.1 Prepare material

Materials

- 48 well plate (CytoOne, #CC7672-7548)
- Tungsten needles (prepared as described by Anastasia R. Nast⁸⁴)
- Parafilm, cut into squares, a little bigger than the diameter of the well (12-13mm length).
- VWR laboratory tape (VWR #89097)
- Elmer's paste or silicone seal (Thermo Fisher Scientific #P18175)
- inDrop device⁴³ (provided by Harvard Medical School Single Cell Core (<https://singlecellcore.hms.harvard.edu/>))
- Micro medical tubing (Scientific Commodities Incorporated #BB31695-PE/2)
- 15 ml falcon tube
- 2ml Pasteur pipettes (VWR #63A54)

- Filtered Artificial Sea Water (FASW) (Artificial SeaWater made from Instant Ocean #671442 filtered with a Nalgene bottle-top 0.2um filter (Sigma #Z358215))
- Sylgard plate: 10 cm Petri dish filled halfway with Sylgard (Sigma Aldrich #761036) or 2% agar in FASW plate: 10cm Petri dish filled halfway with 2% agar in FASW.
- Vortex Adapter for 48 well plate (ThermoFisher #AM10014)

Coat everything that will touch cells with BSA for at least 2 hours as follows:

To coat an object, apply the BSA solution such that any plastic or glass that might come in contact with cells is covered by the solution.

- Prepare 1% BSA solution in ddH₂O.
- Coat 8 wells from 48 well plate with 1%BSA by pouring BSA solution into the wells.
- Coat Pasteur pipettes for embryos and cell transfers with 1%BSA by aspirating the solution in the pipette and leaving it filled.
- Coat Micro medical tubing and syringes by mounting the tube on the syringe and aspirating 1ml of BSA solution.
- Coat 15 ml falcon tube by pouring BSA solution into it.

Prepare 40ml of dissociation buffer

To prepare the dissociation buffer, mix the following ingredients in 40ml of ddH₂O.

- 600mg of Isethionic acid sodium salt (Sigma-Aldrich #220078-25G)
- 360mg of Sodium pyrophosphate tetrabasic decahydrate (Sigma-Aldrich #S6422-100G)
- 88mg of CAPS (3-(Cyclohexylamino)-1-propanesulfonic acid Sigma-Aldrich #C2632-25G)

Prepare 2ml each of 5 different solutions of Optiprep in 1x PBS (All volumes in ul)

	5%	10%	20%	30%	40%
10x PBS	200	200	200	200	200
ddH ₂ O	1695	1600	1400	1195	1000
Phenol red (Sigma-Aldrich #P0290-100ML)	5	0	0	5	0
Optiprep (Sigma-Aldrich #D1556-250ML)	100	200	400	600	800

4.5.2 Collect embryos

This protocol was tested on embryos at 12-24hpf and 60-64hpf.

Collect mated females that are carrying embryos, harvest embryos from the brood pouch, and place them in FASW in a petri dish.

Prepare the dissociation wells

- Remove the 1% BSA from the wells of the 48 well plate and wash with the dissociation buffer once. Then fill halfway (around 3ml) with the dissociation buffer.
- On the lid, locate the position of the wells and place a ring of Elmer's paste that will serve as a seal for each well that will be used. The ring forces the parafilm to stay pressed between the lid and the well, preventing air bubbles from coming in.
- Place the tape on the edges of the lid to allow for it to stay in place once closed.

See Figure 4.21 for a schematic representation of the process.

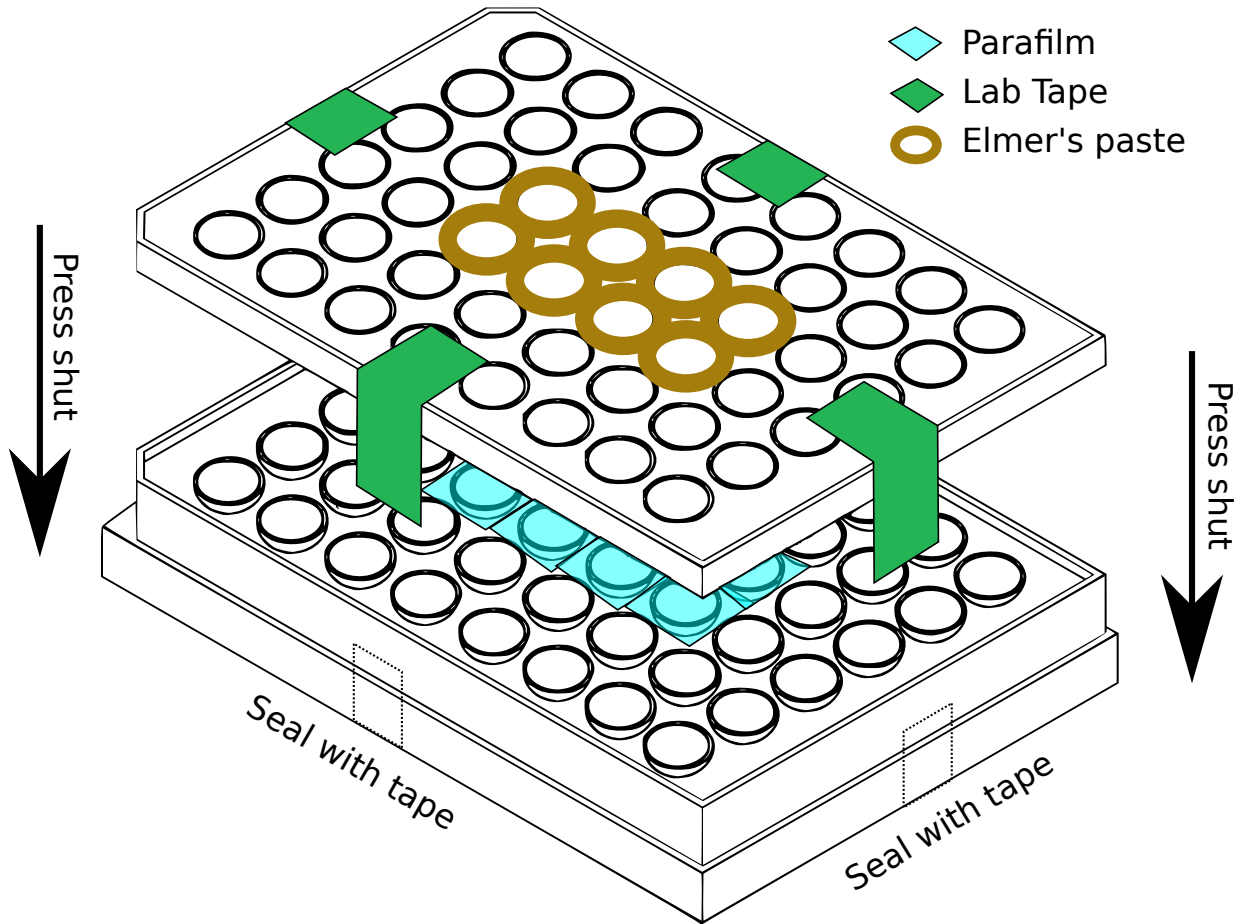


Figure 4.21: Schematic representation of the plate with Elmer's paste seals, parafilm, and lab tape. 48 well plate design adapted from MatTek.

Remove Egg Shell

- Place 160 embryos on a Sylgard plate or a 2% agar in FASW plate in a small drop (around 200-300ul) of FASW.
- Separate the embryos on the plate into 8 pools of 20 embryos per pool, each in a drop of FASW (around 50-100ul).
- One drop at a time, aspirate as much as possible of the FASW away from the embryos.
- Wash with a drop of dissociation buffer by dropping dissociation buffer onto the embryos, then removing the liquid away from the embryos and discarding it. Make sure not to aspirate embryos in the process. Repeat until no white precipitate caused by the buffering of calcium by CAPS is visible.
- Repeat for all 8 drops containing embryos.
- Drop by drop, using the tungsten needles, remove the eggshell from each embryo.
- Make sure to remove the eggshells from the drop once they are removed from the embryos, as if they remain with the embryos in the next step, they will prevent the correct dissociation, and clog the inDrop device.
- Once the eggshell has been removed from all embryos, transfer the entire drop containing the embryos into one of the wells of the 48-well plate, which was previously filled with the dissociation buffer. More than 20 embryos per well results in a failed dissociation (data not shown).

Seal the well for dissociation

- Fill each well containing embryos all the way to the top with the dissociation buffer until a meniscus is visible.
- Place a parafilm square on the meniscus. **THERE MUSN'T BE ANY BUBBLES TRAPPED.**

- Close the lid, sealing the parafilm in place thanks to the paste previously placed on the lid, and fix it with the tape around.

See Figure 4.21 for a schematic representation of the process.

Dissociate the cells

- Place the sealed 48 well plate onto the Vortex plate adapter, and start the vortex at max speed for 20 minutes.

Clean the dissociation buffer

- During the 20 minute dissociation step, wash the 15ml falcon tube filled with BSA with PBS.
- Stop the vortex, remove the plate.
- One well at a time, transfer the contents to the 15ml BSA coated tube. (Figure 4.22 step I)
- Using a syringe with a 21g needle, backfill the tube with Optiprep PBS 60ul 5%, 60ul 10%, 20ul 20%, 20ul 30%, and 400ul 40%. To backfill the tube, start by injecting 60ul of the 5% optiprep solution, then the 60ul of 10% Optiprep solution, all the way to the 40% Optiprep solution. Inject the liquid very slowly to prevent the mixing of the different densities. The tip of the needle must touch the bottom of the falcon tube. (Figure 4.22 step II-V)
- Spin the tube in a bucket centrifuge at 2500rpm for 4 minutes at 4°C. (Figure 4.22 step VI)
- Remove the tube and aspirate carefully above the first red band (5% optiprep) removing all of the dissociation buffer. (Figure 4.22 step VII)
- Then very gently, pipette up and down with the Pasteur pipette the 5%, 10%, 20%, and 30% content that is marked by the two red bands (the 5% layer and the 30% layer). While doing this, be very gentle so that the 40% optiprep layer does

not get disturbed. (Figure 4.22 step VIII)

- This is important because at the bottom of the tube is a lot of debris. That debris if resuspended will interfere with the microfluidic flow.
- Finally, aspirate the mixed Optiprep from 5% to 30% which is red-colored, and backfill the BSA coated syringe. (Figure 4.22 step IX)

Microfluidic flow

- Connect the BSA coated syringe containing the cells to the cell port on an inDrop device and start the flow of cells at 250ul/h.

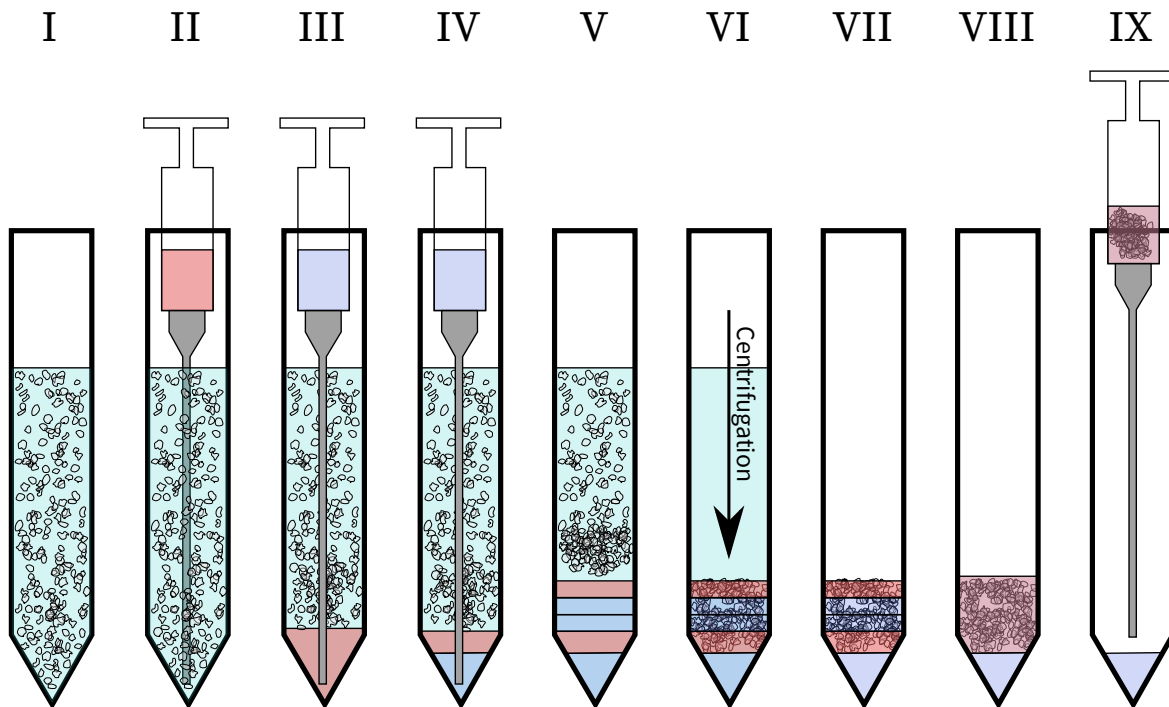


Figure 4.22: Schematic representation of the concentration steps. I) The cells are transferred to the 15ml Falcon tube. II) 5% Optiprep solution is backfilled with a 21G needle. III-V) The different concentrations of Optiprep are backfilled from the lowest density to the highest density. VI) The Falcon tube is centrifuged at 2500 RPM for 4 minutes at 4°C. VII) The dissociation buffer is removed from the falcon tube. VIII) The cells are resuspended and the gradient density is mixed except for the 40% layer. IX) The concentrated cells are aspirated in the syringe pre-coated with BSA.

4.6 List of overrepresented sequences in the two libraries

- ACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTTC
- CGACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- ATGGTTTCTATCTTTTGTTTATATACTCAATAATTTATTTAGTACGAAAGGATTAATAA
- GCCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTTC
- GGGTAGTTTATTTATTTTAATTAATTTTAATTAATACTTAATGGTTTTAAGAGCCTTTAA
- GACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTTC
- TTCTTATTTATATATTTGATTGCGACCTCGATGTTGAATTAATGACTCTTTATAGAGC
- CGTAATAGGTATTAGAAGTTCTTATTTATATATTTGATTGCGACCTCGATGTTGAATTAAT
- AGCCTTTAAATAAAGATTAATAGAAAAAGTTACTTTAGGGCTAACAGCGTAATAGGTATT
- TGACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- GGCTAACAGCGTAATAGGTATTAGAAGTTCTTATTTATATATTTGATTGCGACCTCGATGT
- TCCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTTC
- GCGACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- CCAGATTGGTTTCTATCTTTTGTTTATATACTCAATAATTTATTTAGTACGAAAGGATTA
- ATTGCGACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTT
- CCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTCA
- GTAATTTTTATTTTGTTTTTATATAAATTTATTTAGAAATTTTAACTGGGGTAGTTTATT
- CCCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTTC
- GCCTGCCGATGTAATTATGAATGGCTGCGGTATTGTGACCGTGCTAAGGTAGCATAATC
- TCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTCAACT
- GGGGTAGTTTATTTATTTTAATTAATTTTAATTAATACTTAATGGTTTTAAGAGCCTTTA
- GTTGGTTTCTATCTTTTGTTTATATACTCAATAATTTATTTAGTACGAAAGGATTAATAA
- ATCCTTATTTTATTATTATCTCAGGTTTAATTTTAACATTAATTCTCATTCTTGGTATTAT
- CCCCAGCTGACCATAGATATGATGACACGCCTATTTTAAATAATTAATAAAAAAAAAAAAAA
- GGTAGTTTATTTATTTTAATTAATTTTAATTAATACTTAATGGTTTTAAGAGCCTTTAAT
- GTTCTATCTTTTGTTTATATACTCAATAATTTATTTAGTACGAAAGGATTAATAAATTTT
- ATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTCACTTT
- AGCCAGATTGGTTTCTATCTTTTGTTTATATACTCAATAATTTATTTAGTACGAAAGGATT
- CGGCCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- GGGCTAACAGCGTAATAGGTATTAGAAGTTCTTATTTATATATTTGATTGCGACCTCGATG
- ACCCCGTAAACAAAATTCTATTTTGATCATTGTAGTAATTTTAATTTTATTAACATGGAT

- TGCCCGATGTAATTATGAATGGCTGCGGTATTGTGACCGTGCTAAGGTAGCATAATCATT
- CCCGATGTAATTATGAATGGCTGCGGTATTGTGACCGTGCTAAGGTAGCATAATCATTGT
- GTGGCCAAAAAGTTTTTTTATCAATAAGAAATAAAATTCAAGTAGCTCAGCTAAGAGAAAT
- GCCAGATTGGTTTCTATCTTTTGTATATACTCAATAATTTATTTAGTACGAAAGGATT
- TGCGACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTG
- TGTAATAGGTATTAGAAGTTCTTATTTATATATTTGATTGCGACCTCGATGTTGAATTAAT
- TGGGGTAGTTTATTTATTTAATTAATTTAATTAATACTTAATGGTTTTAAGAGCCTT
- AGGGCTTGTTTTAATCGATAATCCGCGCTTAGTTCTACTTATCTGATTTTTATATATCG
- GCGTTATTCTGATTATCCTGATTCTTACTCTGCTTGAATATAGTTTCTTCTTAGGATCT
- CACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- TACCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- ACCACGATAAATTGCTTAAGTATAAATTTAATAGAGTTAGCTCCTTTAAATATTTATAAGG
- CCTTTAAATAAAGATTAATAGAAAAAGTTACTTTAGGGCTAACAGCGTAATAGGTATT
- CACGGTGTATATTAAGTTAATAGAGGCTTCCATAGAGGGTTTATACTTGAATATCTC
- GGGCTTGTTTTAATCGATAATCCGCGCTTAGTTCTACTTATCTGATTTTTATATATCGTT
- TTCTTTATTTTTATATTATATTATTAACCAATTATAATAAAAATATGAGATTTAGTACC
- TTCAACTTTAAAATTATTACATGATTTGAGTTCAAATCGGTTAAGCCAGATTGGTTTCT
- TTTGGTTTCTATCTTTTGTATATACTCAATAATTTATTTAGTACGAAAGGATTAATAA
- ACCTTGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- CGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTCAACT
- ACCACGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- GCCCGATGTAATTATGAATGGCTGCGGTATTGTGACCGTGCTAAGGTAGCATAATCATTG
- CTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTTCAAC
- TTGGTTTCTATCTTTTGTATATACTCAATAATTTATTTAGTACGAAAGGATTAATAAAT
- ATCTCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- TAGGTGGTTTAAACAGGAGTAATACTTGCTAACTCTTCTATTGATATTATCCTTCATGATAC
- GGCCTTTAAATAAAGATTAATAGAAAAAGTTACTTTAGGGCTAACAGCGTAATAGGTATT
- ACCCGATGTTGAATTAATGACTCTTTATAGAGCAGAATATATAAAAAGAAAGTTTGTT
- ACTGGGGTAGTTTATTTATTTAATTAATTTAATTAATACTTAATGGTTTTAAGAGCCTT
- AGGGCTAACAGCGTAATAGGTATTAGAAGTTCTTATTTATATATTTGATTGCGACCTCGAT
- GCTAACAGCGTAATAGGTATTAGAAGTTCTTATTTATATATTTGATTGCGACCTCGATGTT
- GGCTTGTTTTAATCGATAATCCGCGCTTAGTTCTACTTATCTGATTTTTATATATCGTT

References

- [1] M. Gerberding, W. E. Browne, and N. H. Patel. Cell lineage analysis of the amphipod crustacean *Parhyale hawaiiensis* reveals an early restriction of cell fates. *Development*, 129(24):5789–5801, December 2002.
- [2] C. R. Shoemaker. Observations on the Amphipod Genus *Parhyale*. *Proceedings of the United States National Museum*, 106(3372):345–358, 1956.
- [3] S. Poovachiranon, K. Boto, and N. Duke. Food preference studies and ingestion rate measurements of the mangrove amphipod *Parhyale hawaiiensis* (Dana). *Journal of Experimental Marine Biology and Ecology*, 98(1-2):129–140, 1986.
- [4] D. Kao, A. G. Lai, E. Stamataki, S. Rosic, N. Konstantinides, E. Jarvis, A. Di Donfrancesco, N. Pouchkina-Stancheva, M. Sémon, M. Grillo, H. Bruce, S. Kumar, I. Siwanowicz, A. Le, A. Lemire, M. B. Eisen, C. Extavour, W. E. Browne, C. Wolff, M. Averof, N. H. Patel, P. Sarkies, A. Pavlopoulos, and A. Aboobaker. The genome of the crustacean *Parhyale hawaiiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. *Elife*, 5:e20062, November 2016.
- [5] E. J. Rehm, R. L. Hannibal, R. C. Chaw, M. A. Vargas-Vila, and N. H. Patel. The crustacean *Parhyale hawaiiensis*: a new model for arthropod development. *Cold Spring Harb. Protoc.*, 2009(1):db.em0114, January 2009.
- [6] W. E. Browne, A. L. Price, M. Gerberding, and N. H. Patel. Stages of embryonic development in the amphipod crustacean, *Parhyale hawaiiensis*. *Genesis*, 42(3):124–149, July

2005.

- [7] F. Alwes, B. Hinchén, and C. G. Extavour. Patterns of cell lineage, movement, and migration from germ layer specification to gastrulation in the amphipod crustacean *Parhyale hawaiiensis*. *Dev. Biol.*, 359(1):110–123, November 2011.
- [8] J. M. Serano, A. Martín, D. M. Liubicich, E. Jarvis, H. S. Bruce, K. La, W. E. Browne, J. Grimwood, and N. H. Patel. Comprehensive analysis of Hox gene expression in the amphipod crustacean *Parhyale hawaiiensis*. *Dev. Biol.*, 409(1):297–309, January 2016.
- [9] C. Wolff, J.-Y. Tinevez, T. Pietzsch, E. Stamatakis, B. Harich, L. Guignard, S. Preibisch, S. Shorte, P. J. Keller, P. Tomancak, and A. Pavlopoulos. Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. *Elife*, 7:e34410, March 2018.
- [10] P. Nestorov, F. Battke, M. P. Levesque, and M. Gerberding. The maternal transcriptome of the crustacean *Parhyale hawaiiensis* is inherited asymmetrically to invariant cell lineages of the ectoderm and mesoderm. *PLoS One*, 8(2):e56049, February 2013.
- [11] V. Zeng, K. E. Villanueva, B. S. Ewen-Campen, F. Alwes, W. E. Browne, and C. G. Extavour. De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean *Parhyale hawaiiensis*. *BMC Genomics*, 12:581, November 2011.
- [12] D. A. Sun and N. H. Patel. The amphipod crustacean *Parhyale hawaiiensis*: An emerging comparative model of arthropod development, evolution, and regeneration. *Wiley Interdiscip. Rev. Dev. Biol.*, 8(5):e355, September 2019.
- [13] A. Pavlopoulos, Z. Kontarakis, D. M. Liubicich, J. M. Serano, M. Akam, N. H. Patel, and M. Averof. Probing the evolution of appendage specialization by Hox gene misexpression in an emerging model crustacean. *Proceedings of the National Academy of Sciences*, 106(33):13897–13902, 2009.
- [14] F. Alwes, C. Enjolras, and M. Averof. Live imaging reveals the progenitors and cell dynamics of limb regeneration. *Elife*, 5:e19766, October 2016.

- [15] A. L. Price, M. S. Modrell, R. L. Hannibal, and N. H. Patel. Mesoderm and ectoderm lineages in the crustacean *Parhyale hawaiiensis* display intra-germ layer compensation. *Developmental Biology*, 341(1):256–266, 2010.
- [16] M. Gerberding and N. H. Patel. Gastrulation in Crustacean: Germ Layers and Cell Lineages. In C. D. Stern, editor, *Gastrulation: From Cells to Embryo*, pages 79–90. CSHL Press, 2004.
- [17] H. L. Sladitschek and P. A. Neveu. A gene regulatory network controls the balance between mesendoderm and ectoderm at pluripotency exit. *Mol. Syst. Biol.*, 15(12):e9043, December 2019.
- [18] M. Thomson, S. J. Liu, L.-N. Zou, Z. Smith, A. Meissner, and S. Ramanathan. Pluripotency Factors in Embryonic Stem Cells Regulate Differentiation into Germ Layers. *Cell*, 145(6): 875–889, 2011.
- [19] S. Jang, S. Choubey, L. Furchtgott, L.-N. Zou, A. Doyle, V. Menon, E. B. Loew, A.-R. Krostag, R. A. Martinez, L. Madisen, B. P. Levi, and S. Ramanathan. Dynamics of embryonic stem cell differentiation inferred from single-cell transcriptomics show a series of transitions through discrete cell states. *Elife*, 6:e20487, March 2017.
- [20] N. Wakabayashi-Ito and Y. T. Ip. Mesoderm Formation in the *Drosophila* Embryo. In H. Sink, editor, *Muscle Development in Drosophila*, pages 28–37. Springer New York, New York, NY, 2006.
- [21] L. Alegretti, G. de Aragão Umbuzeiro, and M. N. Flynn. Biologia populacional de *Parhyale hawaiiensis* associada ao fital, Itanhém, São Paulo. *RevInter*, 8(3), 2015.
- [22] L. Alegretti, G. d. A. Umbuzeiro, and M. N. Flynn. Population Dynamics of *Parhyale Hawaiiensis* (Dana, 1853) (Amphipoda: Hyalidae) Associated with an Intertidal Algal Belt in Southeastern Brazil. *J. Crustacean Biol.*, 36(6):785–791, November 2016.
- [23] W. Tadros and H. D. Lipshitz. The maternal-to-zygotic transition: a play in two acts. *Development*, 136(18):3033–3042, September 2009.

- [24] A. Ephrussi, L. K. Dickinson, and R. Lehmann. Oskar organizes the germ plasm and directs localization of the posterior determinant nanos. *Cell*, 66(1):37–50, July 1991.
- [25] C. H. Waddington. *Organisers and genes*. Cambridge Biological Studies. University Press, Cambridge., 1940.
- [26] T. Hashimshony, F. Wagner, N. Sher, and I. Yanai. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, 2(3):666–673, September 2012.
- [27] J. A. Briggs, C. Weinreb, D. E. Wagner, S. Megason, L. Peshkin, M. W. Kirschner, and A. M. Klein. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, 360(6392), June 2018.
- [28] D. R. Farnsworth, L. Saunders, and A. C. Miller. A Single-Cell Transcriptome Atlas for Zebrafish Development. *Developmental Biology*, 459(2):100–108, March 2020.
- [29] J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, and A. F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392), June 2018.
- [30] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, 33(5):495–502, May 2015.
- [31] T. Trcek, T. Lionnet, H. Shroff, and R. Lehmann. mRNA quantification using single-molecule FISH in *Drosophila* embryos. *Nat. Protoc.*, 12(7):1326–1348, July 2017.
- [32] J. R. Moffitt, J. Hao, G. Wang, K. H. Chen, H. P. Babcock, and X. Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U. S. A.*, 113(39):11046–11051, September 2016.
- [33] J. R. Moffitt, J. Hao, D. Bambah-Mukku, T. Lu, C. Dulac, and X. Zhuang. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. U. S. A.*, 113(50):14456–14461, December 2016.

- [34] J. Hartmann, M. Wong, E. Gallo, and D. Gilmour. An image-based data-driven analysis of cellular architecture in a developing tissue. *Elife*, 9:e55913, June 2020.
- [35] Z. Kontarakis and A. Pavlopoulos. Transgenesis in Non-model Organisms: The Case of *Parhyale*. In Y. Graba and R. Rezsöházy, editors, *Hox Genes: Methods and Protocols*, pages 145–181. Springer New York, New York, NY, 2014.
- [36] E. J. Rehm, R. L. Hannibal, R. C. Chaw, M. A. Vargas-Vila, and N. H. Patel. Fixation and dissection of *Parhyale hawaiiensis* embryos. *Cold Spring Harb. Protoc.*, 2009(1):db.prot5127, January 2009.
- [37] *Drosophila* apple juice-agar plates. *Cold Spring Harb. Protoc.*, 2011(9):db.rec065672, September 2011.
- [38] W. F. Rothwell and W. Sullivan. *Drosophila* embryo collection. *CSH Protoc.*, 2007: db.prot4825, September 2007.
- [39] E. J. Rehm, R. L. Hannibal, R. C. Chaw, M. A. Vargas-Vila, and N. H. Patel. Antibody staining of *Parhyale hawaiiensis* embryos. *Cold Spring Harb. Protoc.*, 2009(1):db.prot5129, January 2009.
- [40] R. G. Martinho, P. S. Kunwar, J. Casanova, and R. Lehmann. A noncoding RNA is required for the repression of RNAPolIII-dependent transcription in primordial germ cells. *Curr. Biol.*, 14(2):159–165, January 2004.
- [41] C. G. Extavour. The fate of isolated blastomeres with respect to germ cell formation in the amphipod crustacean *Parhyale hawaiiensis*. *Dev. Biol.*, 277(2):387–402, January 2005.
- [42] T. Hashimshony, N. Senderovich, G. Avital, A. Klochandler, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, Y. Dor, A. Regev, and I. Yanai. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, 17:77, April 2016.
- [43] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.

- [44] C. Holt and M. Yandell. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12:491, December 2011.
- [45] J. M. Crescente, D. Zavallo, M. Helguera, and L. S. Vanzetti. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics*, 19(1):348, October 2018.
- [46] D. Ellinghaus, S. Kurtz, and U. Willhoeft. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9:18, January 2008.
- [47] J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.*, 117(17):9451–9457, April 2020.
- [48] N. S. Vassetzky and D. A. Kramerov. SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res.*, 41(Database issue):D83–9, January 2013.
- [49] SMIT and F. A. A. Repeat-Masker Open-3.0. <http://www.repeatmasker.org>, 2004.
- [50] W. Bao, K. K. Kojima, and O. Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, 6:11, June 2015.
- [51] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, 37(8):907–915, August 2019.
- [52] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.
- [53] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, 33(3):290–295, March 2015.

- [54] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, 7(3):562–578, March 2012.
- [55] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, 34(Web Server issue):W435–9, July 2006.
- [56] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, October 2015.
- [57] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O’Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, 31(1):365–370, January 2003.
- [58] F. A. Wolf, P. Angerer, and F. J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, February 2018.
- [59] B. Bartkowiak, A. L. Mackellar, and A. L. Greenleaf. Updating the CTD Story: From Tail to Epic. *Genet. Res. Int.*, 2011:623718, October 2011.
- [60] A. Pavlopoulos and M. Averof. Establishing genetic transformation for comparative developmental studies in the crustacean *Parhyale hawaiiensis*. *Proc. Natl. Acad. Sci. U. S. A.*, 102(22):7888–7893, May 2005.
- [61] B. M. Gumbiner. Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell*, 84(3):345–357, February 1996.
- [62] S. C. van den Brink, F. Sage, Á. Vértesy, B. Spanjaard, J. Peterson-Maduro, C. S. Baron, C. Robin, and A. van Oudenaarden. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods*, 14(10):935–936, September 2017.
- [63] F. R. Horne. The Effect of Digestive Enzymes on the Hatchability of *Artemia salina* Eggs. *Transactions of the American Microscopical Society*, 85(2):271, 1966.

- [64] S. Andrews and Others. FastQC: a quality control tool for high throughput sequence data. 2010.
- [65] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, September 1997.
- [66] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezuk, S. McGinnis, and T. L. Madden. NCBI BLAST: a better web interface. *Nucleic Acids Res.*, 36(Web Server issue):W5–9, July 2008.
- [67] E. Keogh and A. Mueen. Curse of Dimensionality. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*, pages 314–315. Springer US, Boston, MA, 2017.
- [68] C. Weinreb, S. Wolock, and A. M. Klein. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(7):1246–1248, April 2018.
- [69] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, 9(1):5233, March 2019.
- [70] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. February 2018.
- [71] V. Ntranos, L. Yi, P. Melsted, and L. Pachter. Identification of transcriptional signatures for cell types from single-cell RNA-Seq. *bioRxiv*, page 258566, February 2018.
- [72] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47(D1):D506–D515, January 2019.
- [73] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, 47(D1):D330–D338, January 2019.
- [74] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene

- ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000.
- [75] M. A. Vargas-Vila, R. L. Hannibal, R. J. Parchem, P. Z. Liu, and N. H. Patel. A prominent requirement for *single-minded* and the ventral midline in patterning the dorsoventral axis of the crustacean *Parhyale hawaiiensis*. *Development*, 137(20):3469–3476, October 2010.
- [76] F. Kilchherr, S. Baumgartner, D. Bopp, E. Frei, and M. Noll. Isolation of the paired gene of *Drosophila* and its spatial expression during early embryogenesis. *Nature*, 321(6069):493–499, May 1986.
- [77] A. Preiss, D. A. Hartley, and S. Artavanis-Tsakonas. The molecular genetics of *Enhancer of split*, a gene required for embryonic neural development in *Drosophila*. *EMBO J.*, 7(12):3917–3927, December 1988.
- [78] M. Whiteley, P. D. Noguchi, S. M. Sensabaugh, W. F. Odenwald, and J. A. Kassis. The *Drosophila* gene *escargot* encodes a zinc finger motif found in snail-related genes. *Mech. Dev.*, 36(3):117–127, February 1992.
- [79] O. Johnstone, R. Deuring, R. Bock, P. Linder, M. T. Fuller, and P. Lasko. Belle is a *Drosophila* DEAD-box protein required for viability and in the germ line. *Dev. Biol.*, 277(1):92–101, January 2005.
- [80] N. H. Putnam, B. L. O’Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J. Troll, A. Fields, P. D. Hartley, C. W. Sugnet, D. Haussler, D. S. Rokhsar, and R. E. Green. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.*, 26(3):342–350, March 2016.
- [81] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.
- [82] A. A. Hardigan, B. S. Roberts, D. E. Moore, R. C. Ramaker, A. L. Jones, and R. M. Myers. CRISPR/Cas9-targeted removal of unwanted sequences from small-RNA sequencing libraries. *Nucleic Acids Res.*, 47(14):e84, August 2019.

- [83] W. Gu, E. D. Crawford, B. D. O'Donovan, M. R. Wilson, E. D. Chow, H. Retallack, and J. L. DeRisi. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.*, 17:41, March 2016.
- [84] C. G. E. Anastasia R. Nast. Ablation of a Single Cell From Eight-cell Embryos of the Amphipod Crustacean *Parhyale hawaiiensis*. *J. Vis. Exp.*, (85), 2014.

The scientist does not study nature because it is useful to do so. He studies it because he takes pleasure in it, and he takes pleasure in it because it is beautiful. If nature were not beautiful it would not be worth knowing, and life would not be worth living.

Henri Poincaré, 1908

5

Studying germ layer specification in the early embryogenesis of *Parhyale hawaiiensis* with modern microscopy

ABSTRACT

Gastrulation is a fundamental morphogenetic event that often orchestrates the specification of germ layers. In this chapter, I used recent advances in light sheet microscopy to record *in toto* the first four days of *Parhyale hawaiiensis* embryogenesis. To examine the early germ layer specification of *Parhyale hawaiiensis*, I tracked each of the lineages. By establishing a ground truth annotation of an embryo by tracking every cell I showed a correlation between developmental stages, division rate, and nuclei velocity. Moreover, I provide novel insights into the dynamics of each germ layer's contribution to the total cell population in the embryo. Finally, I ablated the ectodermal Er blastomere in four embryos to study the changes in cellular behavior happening during the regeneration of the missing lineage. I successfully recorded the first three days following ablation using light sheet microscopy. Finally, I generated valuable resources for the study of the early development of *Parhyale hawaiiensis*.

CONTRIBUTIONS

For this study, I received the help of multiple members from Philip Keller's laboratory, in particular, Bill Lemon who assisted and taught me how to manipulate the SIMView microscope. Anastasios Pavlopoulos and Evangelia Stamataki funded in part this study and invited me to come to Janelia Farm to perform those experiments. Moreover, they performed the injection of the 2017 datasets, taught me how to inject mRNA into one cell stage embryos and provided the injected mRNAs. Cassandra G. Extavour assisted and taught me how to ablate single blastomeres of *Parhyale hawaiiensis* embryos.

ACKNOWLEDGEMENT

This study would not have been possible without the help I received from Anastasios (Tassos) Pavlopoulos and Evangelia (Valia) Stamataki. They provided us with the *P. hawaiiensis* lines used throughout this thesis, invited me and Cassandra Extavour to the HHMI Janelia Research Campus, and taught me how to produce mRNA and inject it into *P. hawaiiensis* embryos. Moreover, I also want to thank A. Pavlopoulos for the long discussions on the biology of crustaceans and his willingness to help and answer any question. I want to thank Cassandra Extavour for supporting the ideas I attempted here, and her help in teaching me microdissection and ablation techniques. Finally, I want to thank the members of Philip Keller's Laboratory who helped me with the complex microscopes they developed, especially Bill Lemon and Kate McDole who spent hours teaching me the intricacies of their machines.

5.1 Introduction

In this chapter, I present my advances towards the recording of live *P. hawaiiensis* embryos. I then present the results of the tracking of nuclei in one of the recorded embryos. Finally, I conclude chapters 4 and 5 with a general discussion regarding the potential for a future generation of any embryonic developmental atlas.

5.1.1 The uses of Selective Plane Illumination Microscopy to study live developing embryos

In 1903, Siedentopf and Zsigmondy set out to measure the size of gold nanoparticles in rubies¹. Because the size of the particles was smaller than the wavelength of visible light, classical light absorption microscopy did not work¹. Using the work of Lord Rayleigh (John William Strutt) on the scattering of light by small particles², they used Rayleigh scattering to study the size of the gold nanoparticles. To this end, they invented the ultramicroscope, which shines condensed light on a sample at a high angle from the observing objective¹. In this way, only light scattered by the particles would reach the observing objective, and all remaining light would pass through the sample¹. By measuring the diffraction ring caused by the scattered light, Siedentopf and Zsigmondy calculated the size of the nanoparticles¹. For this work, they were awarded the 1925 Nobel prize³. This seminal research led to what is known today as darkfield microscopy (discussed in Keller and Dodt⁴), a technique used to observe the scattering of light by a sample from incident illumination created by a ring filter on the path of the condenser (reviewed by Sheppard⁵). The ring prevents direct light ray trajectories from reaching the objective, such that only light scattered by the sample reaches the objective (reviewed by Sheppard⁵). It allows the observation of unstained transparent samples due to the differences in the refractive index of different parts of tissues and cells (reviewed by Sheppard⁵). Example darkfield micrographs are shown in Figure 5.1.

In 2004 Huisken and Stelzer proposed a new microscopic method to study live biological specimens that they called Selective Plane Illumination Microscopy (SPIM) and that became more widely known as light sheet microscopy⁶. The high angle of incidence of the emission

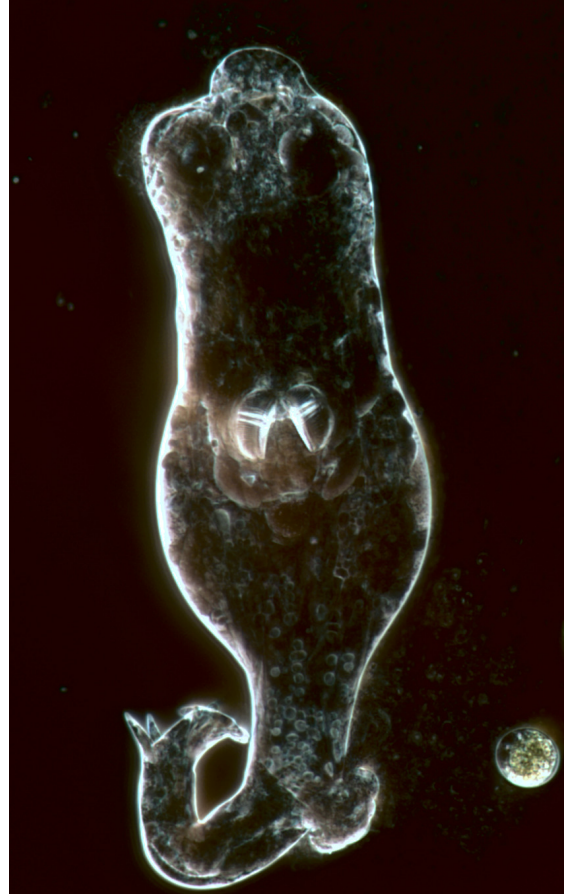
a**b**

Figure 5.1: Example micrographs showing the contrast generated by darkfield microscopy. a) Unknown species of diatoms laid between glass and coverslip. The ultrastructures of the diatoms become visible through this contrast technique. **b)** Example of a live unknown species of *Stentor* (ciliate) found in a river sample from the Charles River in Cambridge, MA. The internal calcified teeth-like structures are visible, as well as some of the internal structures of the animal.

light was akin to dark field microscopy, but the use of fluorescence made it different in that it did not use the scattering of light, but instead the emission of photons from an excited fluorophore⁶. In a SPIM light sheet microscope, an emission laser source is bent into a plane with a thickness of a few micrometers⁶. This sheet can then selectively be moved through a three-dimensional sample embedded in a chamber⁶. The sheet is aimed at 90° from the collection objective, creating an illuminated plane perpendicular to the collection objective path⁶. This achieves a planar excitation of the fluorophores, guaranteeing that only molecules within the focal plane are excited, resulting in very high contrast⁶. This microscopy technique offers multiple advantages, from a low required excitation intensity and therefore low biological damage, with the ability to collect, at higher speeds than before⁷, three-dimensional images of live biological samples (reviewed by Keller et al.⁸). By changing the plane to a vertically scanned laser beam with structured pulses, and timing those pulses to the reading of the corresponding pixel line of a CMOS sensor, Phillip Keller was able to increase the contrast and resolution of deep tissue imaging⁹, paving the way for the invention of the higher speed, higher resolution, and higher contrast (relative to classical SPIM microscopes⁶) SIMView microscope that I used during my Ph.D.⁷.

Light sheet microscopy has been used to study the development of embryos^{6,7,9,10,11,12,13,14}. An example of the advantages of light sheet microscopy in recording embryogenesis is the difficult task of imaging mammalian embryos. Mammalian embryos are particularly sensitive to light and require a specific medium to develop properly outside of a uterus¹⁰. Using light sheet microscopy, Kate McDole successfully recorded the first days of embryogenesis of a mouse embryo at single-cell resolution¹⁰. Isomorphic three-dimensional recordings also allow for the study of morphogenesis. An example is the recent study by Anastasios Pavlopoulos of the developing limb of *P. hawaiiensis*¹¹. In this study, they tracked cells during limb development and found that the resulting cellular architecture of the developing limb could be traced back to the position of the cells in the grid stage¹¹. The high recording speed of light sheet microscopy allowed William C. Lemon to image the activity of neurons in a developing *Drosophila melanogaster* larva. By recording the neuronal activity of the entire larva, he and his colleagues modeled the relation between specific neurons and the backward and forward activation

waves¹⁵. In this thesis, I used this microscopy technique to record the early development of *P. hawaiiensis* under normal conditions, as well as after the ablation of one of the ectodermal blastomeres at the S4 stage.

5.1.2 The advantages and challenges of volumetric microscopy data

There are multiple computational challenges to process and analyze light sheet microscopy datasets. First, the accurate registration of the recording of a sample from multiple angles in time is necessary to correctly assign the position of each recorded image in three-dimensional space. Two main techniques have been used to achieve this. First, the position of the objectives, the cameras, and the sample are fixed, and the recording is done through two⁹ or four objectives¹⁶. This way, it becomes trivial to register the datasets, as the rotation and translation are built into the microscope^{9,16}. However, this technique only allows for at most four angles and is therefore not as accurate as the second method. The second method uses the embedding of microscopic fluorescent beads within the sample embedding medium. Given that the size of the beads is smaller than the wavelength of light, the resulting image is equivalent to a point spread function, allowing for the inference of the position at a higher resolution than that permitted by traditional light microscopy¹⁷. The positions of beads are extracted from each view and cross-correlated through the RANSAC algorithm to extract the rotation, translation, and scale matrices required to realign each sample¹⁷. The positions of the beads can also be used through time to register each image with the others, allowing for the correction of drift¹⁷. As we will see below, I decided to use the second methodology to register the datasets I collected during my Ph.D., as this allowed me to record the embryos from six or eight angles, and to mount and record multiple embryos in parallel.

The second challenge regards the correct generation of a volumetric image (a three-dimensional image composed of voxels) from single two-dimensional planes. This has seen two main developments. The first one was the generation of voxel data through the fusion of the corresponding pixels of the different imaged planes. To compute the voxel value, the weighted average of the multiple planes is computed^{9,17,18}. The second one used a Bayesian deconvolution of the datasets by using the fact that the same location can be observed from

multiple angles (multiple sampling of the same voxel), and that the beads provide an accurate point spread function (giving an estimation of the scattering of light for a sample)¹⁹. While the deconvolution method proposed by its developers is optimized for large three-dimensional datasets, it remains computationally expensive and might generate artifacts¹⁹. Therefore, the volumetric images in this chapter were generated using the weighted average fusion algorithm.

The third challenge regards the segmentation and tracking of objects in three dimensions, which is a very active and non-trivial research topic (reviewed and benchmarked in Ulman et al.²⁰). To this day, this remains a major problem to analyze large three-dimensional datasets containing a large number of cells (reviewed and benchmarked in Ulman et al.²⁰). While tools have been generated to automatically segment, track and analyze large datasets in two dimensions^{21,22,23}, fewer algorithms have been developed for use in three dimensions (reviewed and benchmarked in Ulman et al.²⁰). The first class of algorithms, such as Ilastik²², tries to use the advances in 2D segmentation by analyzing each plane of the volume as a 2D image. But those algorithms usually perform poorly due to the lack of three-dimensional integration (benchmarked in Ulman et al.²⁰). The second class of algorithms fully uses all three dimensions of the dataset, resulting in much heavier computations. One such algorithm is TGMM^{10,24} which finds ellipsoidal objects (such as nuclei) by fitting a 3D gaussian mixture model onto each volume. Other algorithms use recent advances in deep learning to segment volumetric images, such as the StarDist algorithm²⁵ which first classifies pixels using a U-Net²⁶ or ResNet²⁷ architecture by creating two maps, an object center, and a star convex polygon distance map. Another approach is used by CellPose (also based on a U-Net²⁶ architecture) which first learns to segment nuclei on two-dimensional images, and then applies this segmentation across all slices from all three axes, resulting in a 3D probability matrix²⁸.

The final challenge comes from the amount of data generated by light sheet microscopy, where the number of voxels increases as the cube of the image size. For example, a single time point of the fused dataset generated below represents about 2Gb of data, with a whole dataset of 367 time points using 722Gb after processing (prior to processing each dataset represents between 3 and 5 Tb of images). In comparison, the entire *P. hawaiiensis* genome contains 3.6Gb of data²⁹. I believe that the large amount of data and the processing complexity of volumetric data can

make the use of light sheet datasets difficult for non-computer scientists. Thankfully, the increasing computer power and the progress in developing user-friendly plugins for FIJI^{11,17,18,19,25,30} are making such tasks more amenable to non-computer scientists.

To understand the developmental dynamics of *P. hawaiiensis* early embryogenesis, I needed to segment and track the position of nuclei to analyze the behavior of all cells in the embryo. While seminal papers on *P. hawaiiensis* embryogenesis laid the foundation for the understanding of its early embryogenesis^{31,32,33,34,35}, no comprehensive study of the cells in developing embryos from the eight cell stage to the germ band elongation stage had previously been attempted. Finally, it is known that the fate of cells is restricted from the eight cell stage onward, when the blastomeres acquire their identity, and give rise to populations of cells that are separated by clear boundaries. For example, the Ep blastomere gives rise to a population of cells at the germ band stage located at the posterior of the embryo^{31,32,33,34,35}. This population of cells is found at later stages forming the ventral midline between the descendant of E1 and Er^{31,32,33,34,35}, raising a number of unanswered questions: What is the cellular motions and rearrangements necessary for that population of cells to come to be located at the midline? Is the boundary between the right and left ectoderm absolute? What are the intra-embryonic variability in cell behavior, motion, and final fate? I aimed to generate datasets that would be capable of providing initial answers to these questions.

5.2 Methods

5.2.1 Preparation of low melt agar gel with fluorescent beads

This part of the protocol was performed by me. First, a 20ml solution of a 2% low melt agar in FASW was prepared and left in a 65°C incubation oven overnight. To achieve a uniform melting of the agar and avoid artifacts due to light bending on unmelted agar granules, it is critical to leave the agar at 65°C overnight. I prepared a 1:100 dilution of the fluorescent beads (Etapore Microsphere F-Y cat nb: 80380495, initial concentration 1% of beads, dilution is at 0.01%) by vortexing the bead stock for 2 minutes, then diluting 10ul of the stock solution in 990ul of FASW. The final concentration of a stock solution is at 0.0001%. This stock solution was sonicated with

a water bath sonicator (BRANSON 1510) for 10 minutes before any use and kept at 4°C until use. The final mounting medium, 1% low melt agar in FASW with beads, was prepared by mixing 40ul of the sonicated bead stock with 460ul of FASW and 500ul of the 2% low melt agar stock. The solution was vortexed thoroughly for at least 1 minute, then placed at 42°C. The final beads concentration used for imaging is at $4.10^{-6}\%$.

5.2.2 Preparation of mRNA injection mix

This part of the protocol was performed by A. Pavlopoulos and E. Stamataki (Janelia Research Campus). mRNA of fusion proteins for nuclei and membrane tagging was prepared using the mMMESSAGE mMACHINE kit (ThermoFisher: AM1344). Two different mRNA expression plasmids (designed by A. Pavlopoulos) were used: H2A-mCherry (PCS2 + H2A-mCherry:pDest) and Lyn-GFP (PCS2 + Lyn-GFP). The concentration of mRNA was measured, then the mRNA was stored at -20°C in isopropanol before subsequent use.

The injection mix consisted of phenol red dye at 0.13% (Sigma P0290) along with the prepared mRNA at a final concentration of 2ug/ul. A total of 5ul of injection mix was prepared before each injection. The mix was then spun at 13000rpm for 30 min at 4°C before being loaded into a needle (Eppendorf FemtoTip II #5242957000) and kept on ice at all times.

5.2.3 Injection of probe mRNA into one-cell stage embryos

This part of the protocol was performed by E. Stamataki (Janelia Research Campus) for WT01_11-17 and WT02_11-17. I performed all injections for the 2018 wild type datasets and ablated embryos. A. Pavlopoulos and E. Stamataki performed part of the embryo collection. One cell stage embryos (see: Collection of *P. hawaiiensis* embryos) were placed on an injection stand apparatus molded from agar. To create the injection mold (see schematic in Figure 5.2), two glass slides were used to create a small stair shape by placing one on top of the other with a slight offset, and taping the slides into place with lab tape. The taped slides were placed onto a 10cm petri dish. A 1% agar in FASW solution was prepared and poured into the petri dish until the agar reached the top of the glass slides. After the agar gelified, the glass slides were carefully removed. Finally, FASWA was poured onto the petri dish and the embryos were placed onto the

created steps.

Injection of the one-cell stage embryos was performed using a microinjection setup composed of a Narishige IM-300 Microinjector, a three axis micromanipulator, a foot pedal for injection, and Femtotip II needles (Eppendorf: ep5242957000). The needle was backfilled with the injection mix, and visual confirmation of the injection in the embryo was performed using the phenol red dye. After the injection, the embryos were collected and placed in a clean petri dish that was placed in a 26°C incubator.

Starting at two hours following the injection, embryos were checked for expression of Lyn-GFP and H2A-mCherry using a fluorescent dissection microscope. Embryos that died or did not develop properly (e.g. missing cell, arrest in division, etc.) were removed from the petri dish to avoid microbial growth that could contaminate the remaining embryos.

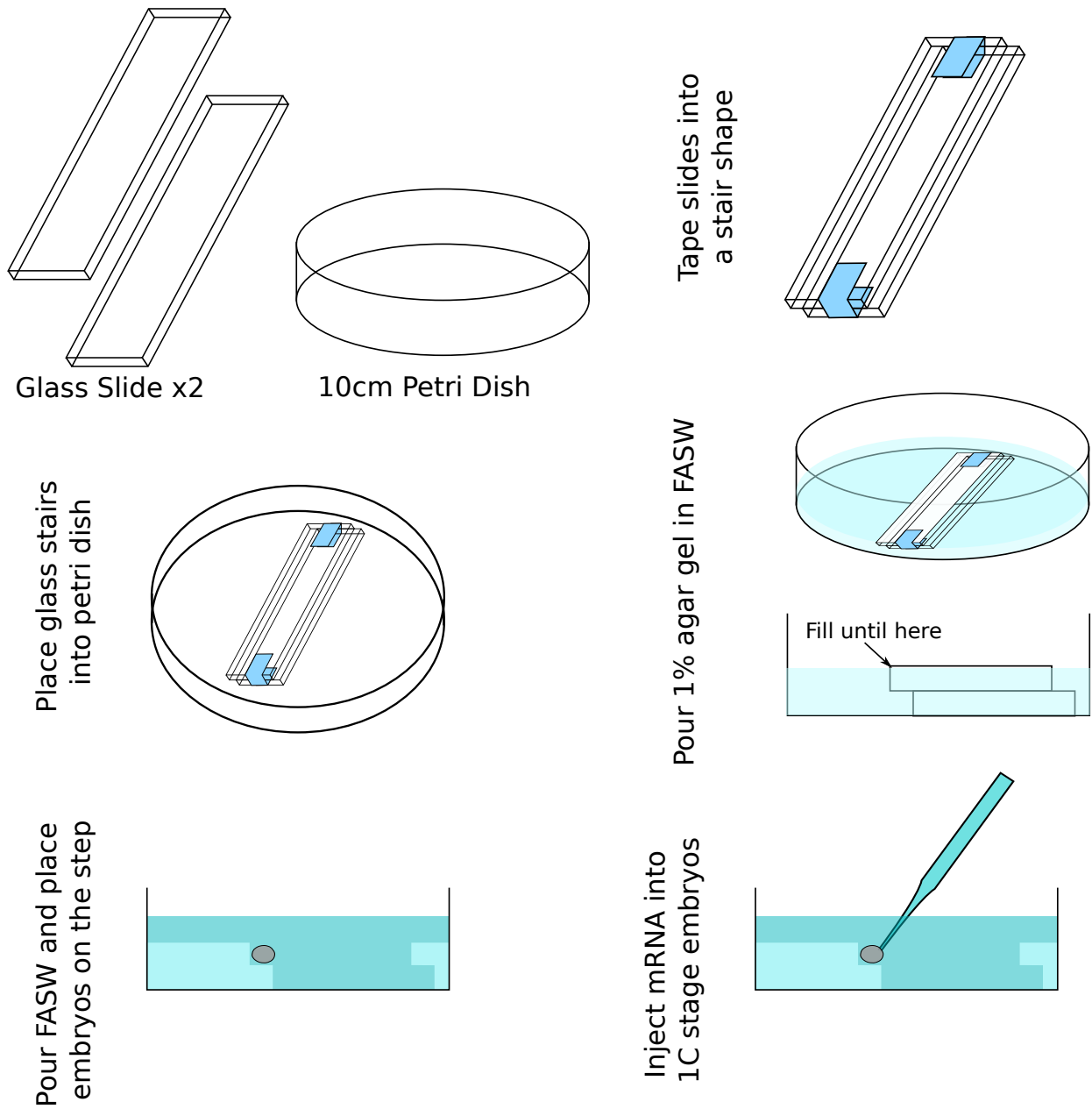


Figure 5.2: Schematic representation of the creation of an agar step mold for injection of *P. hawaiensis* embryos. Glass slides are affixed with tape in a stair shape. Then agar is poured into a petri dish containing the glass slides mold until the agar level is just under the glass slide. The glass slide mold is gently removed taking care to not damage the agar. Finally, the petri dish is filled with FASWA and embryos can be injected atop the first step.

5.2.4 Preparation of embryos for light sheet microscopy with the SimView Keller laboratory scope

Injected embryos were selected for mounting and imaging based on the following criteria, which were assessed through visual inspection on a fluorescent dissection microscope: The embryo must have just reached the four cell stages, and have a strong uniform expression of the injected H2A-mCherry (nuclei) and Lyn-GFP (membrane). The best looking three embryos according to these criteria were then mounted into a PTFE tube (External diameter 2mm, thickness 25um) inserted in a glass capillary (external diameter 3mm, internal diameter 2mm). Both the PTFE tube and the glass capillary were custom made by the Keller laboratory (Janelia Research Campus) to be mounted on the SimView scope. The mounting procedure, shown in Figure 5.3, consisted of creating a plasticine mount onto which a glass capillary was inserted. Inside the capillary, using clean forceps, the PTFE tube is gently pushed, until it reaches the plasticine. Finally, the PTFE tube is backfilled with the 1% low melt agar in FASW with beads until the agar reaches about 1mm above the glass capillary (Figure 5.3 step I). A 10ul fine pipette tip was cut at the end such that the opening was big enough to aspirate an embryo. One embryo from the petri dish (Figure 5.3 step II) was transferred onto the 1% low melt agar with beads solution using the tip and a P10 pipette. The embryo was mixed thoroughly in the low melt agar by swirling the solution around the embryo with the pipette tip (Figure 5.3 step III). Finally, the embryo was aspirated and gently deposited on top of the agar column in the PTFE tube, taking great care to center it in the column (Figure 5.3 step VI). This procedure was repeated until three embryos were mounted on top of each other in the column.

SimView2 Embryo mounting procedure

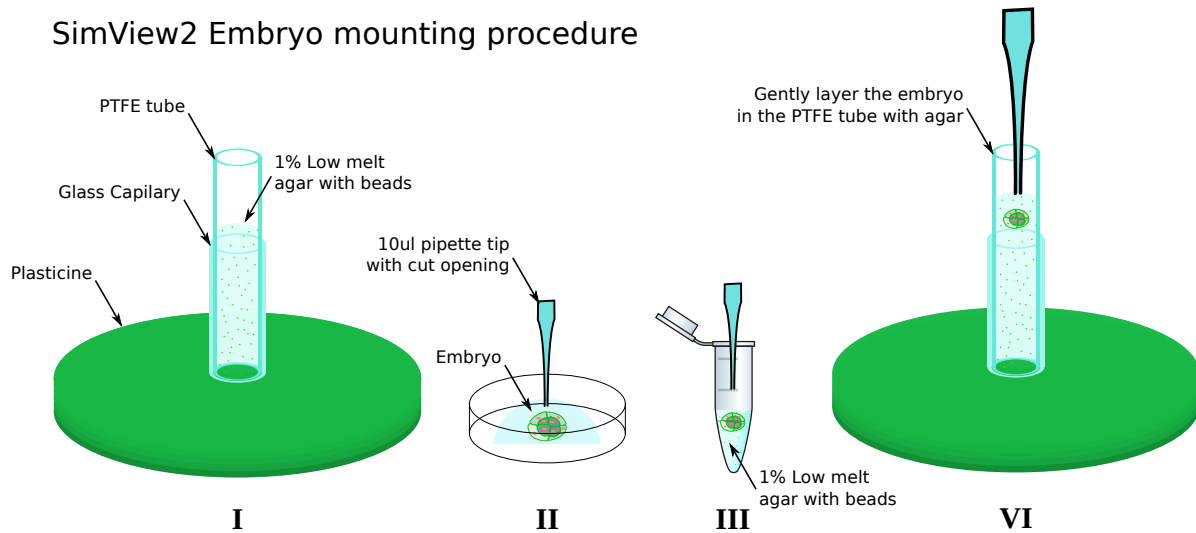


Figure 5.3: Custom mounting procedure for multiple *P. hawaiensis* embryos on the SimView scope. I) A plasticine mount is laid out on a flat surface. A glass capillary is inserted inside the plasticine. Then using forceps, a PTFE tube is inserted into the glass capillary, then filled up until the top of the glass capillary with 1% Low melt agar with fluorescent beads. II) The embryo is selected and picked from a petri dish. III) The embryo is dropped and mixed in liquid 42°C 1% low melt agar with fluorescent beads. IV) The embryo is aspirated with 10ul of agar and laid atop the already solidified agar gel in the column. Extra care is taken to centering it in the tube by gently moving it while the agar is still liquid. IV is then repeated for each new embryo, making sure that each of them is mounted atop each other.

5.2.5 Blastomere ablation

This part of the protocol was performed by myself and C. Extavour at the Janelia Research Campus. Embryos injected with tracer mRNA were selected as previously described (see: Preparation of embryos for light sheet microscopy with the SimView Keller laboratory scope) for ablation. Once the embryos reached the eight cell stage, one blastomere per embryo was ablated using the previously published method³³. The Er blastomere was targeted for ablation each time. However, some embryos had El or Mav removed by mistake instead, which became evident after ablation. Ablated embryos were left to recover for an hour at 26°C before being visually inspected. Three healthy embryos were then selected for mounting as previously described.

5.2.6 Recording of embryos with the SimView microscope

This and all subsequent parts of the protocol were performed by me. The mounted embryos were imaged using the custom microscope SimView at the Keller laboratory in Janelia Research

Campus⁷. Each embryo was imaged using two channels, one for the Lyn-GFP membrane tag using a 488nm emission laser and one for the H2A-mCherry using a 555nm emission laser. The exact settings for each embryo are reported in the Table 5.1. Embryos were imaged with three or four angles, and two objectives located 180° from each other, resulting in six- or eight-angle datasets. The angle spacing was uniform (for example, for 3 angles, spacing was at 0, 60, and 120°, with the corresponding opposite objective recording 180, 240, and 320°). The time resolution was 10mn between time points, except for Ablated-03-06 which was 12mn, which was the minimum amount of time required to complete all scans for four embryos.

5.2.7 Processing and fusion of multiview microphotographs

The files generated by the SimView scopes are KLB formats, one KLB file for each view (a view is defined as one Z-scan for one camera and one angle). The processing was done using the open-source toolkit MultiView Reconstruction, a FIJI plugin¹⁷. This plugin relies on the BigDataViewer FIJI plugin, which allows for the efficient loading and processing of large (multiple terabytes) microscopy datasets³⁰. The dataset metadata was defined into an XML file using the KLB importer tool of BigDataViewer. This XML file is the central file that will be used throughout the whole process. It contains all the necessary information for processing, opening, registration, and fusing. Due to the SimView dual-camera setup, one of the settings of the XML file was manually changed; namely, each dataset captured by the second camera was flipped symmetrically on its Z-axis. To do this, the ViewSetup transform matrix for Angle 1 views was changed, with the m11 value changed from 1 to -1.

Next, using the embedded fluorescent beads, the different views were registered to each other such that each voxel from each view would be correctly aligned for fusion. The registration was performed using the MultiView Reconstruction plugin with the following settings:

- Interest point detection: Channel 0, difference-of-gaussian (sigma=1.7065, threshold=9.701738E-4). All other settings were default values.
- Remove detection by distance: Set to Distance threshold with Min 0 and Max determined as the highest point of the first peak on the shown histogram

composed of two mixed Gaussians (in the range of 10-30px). All other settings were default values.

- Registration: Fast Descriptor based, All to All time point matching with range (10 time points), Only compare overlapping views. All other settings were default values.
- Fusion: Select the bounding box corresponding to the embryo minimizing extraneous space. Make sure to verify across all time points for embryo movement outside the bounding box. Select intensity normalization to global maxima. Select One file per time point. All other settings were default values.

The output created was a collection of H5 (HDF5) files containing the fused dataset, with one file per time point, one H5 file containing metadata about the other H5 files, and one XML file containing the resolution, angles, and location of the H5 data files.

5.2.8 Automatic tracking of nuclei with Ilastik

On the first dataset (WT_01-11-17), nuclei were segmented and tracked using Ilastik, a Random Forest pixel classification machine learning tool²².

Because of the high variability in nuclei intensity throughout the recording time (which I hypothesize was caused by variability between nuclei in the dynamics of the fluorescent protein being expressed, then slowly bleached while not being replaced), the dataset was split into three parts, time points 0-100, 101-200, and 201-300. A model was created for each subset by painting nuclei and background areas. For the training, one time point for every 30 recorded time points was sampled and all nuclei were painted within the volume. Finally, the three subsets were fully segmented.

Then, using the segmented results, the Ilastik Object tracking algorithm was used to track nuclei in all three subsets, merged into a single dataset. Finally, the results were exported using the MaMuT export plugin that is available in Ilastik²² by default, using its default parameters.

5.2.9 Manual correction and tracking of nuclei with MaMuT

Tracks were imported into the Open Source FIJI light sheet nuclei tracking plugin, MaMuT¹¹ for visualization, verification, and correction. The tracks were manually corrected for errors due to the automatic object tracking of Ilastik. Due to the high error rate induced by Ilastik, every single track was manually corrected. The correction was performed as follows: Starting at time point 0 each track was followed one time point at a time. At each division, each track was split and one daughter cell was followed until the next division. When the next division occurred, the track was followed backward to the previous division, and the other track (second daughter cell) was followed. This was continued with one of the two daughter cells until another division happened. The other daughter's cell was then followed until the following division. This process was repeated until all tracks had been manually verified and corrected. The tracks were saved throughout the process, and a final XML file was generated when all tracks were completed. The WT_01-11-17 embryo was tracked in this way up until time point 300 (out of 367).

5.2.10 Analysis of the number of nuclei and rate of division

A custom Python script was created to import the final XML file for subsequent analysis (ref: Chapter_4/Scripts/light sheetUtils/mamut.py). This script imports the tracks into a NetworkX³⁶ Directed Graph (DiGraph) instance for further processing. Using the DiGraph instance, any node (nuclei) that had an out-degree (the number of connecting annotations on the next time point) of two was counted and annotated as a cell division. All division events were stored and their times into an array and subsequently binned them into an equally spaced histogram with a bin size of one hour. Finally, the number of divisions was normalized by the total number of nuclei during the corresponding hour. This computation is the hourly rate of division, which was then plotted using SeaBorn³⁷, a python plotting library.

The total number of nuclei for each frame was computed by counting the number of nodes at each time point. A linear regression of the rate of division was performed for different phases of embryonic development. The following four bins were created: 16-64 cell stage, 64 to 128 cell stage, gastrulation, germ disk formation, and aggregation. For each bin, the division rate was

computed using the Scipy stats linear regression module. The number of nuclei and rate were then plotted using SeaBorn³⁷.

5.2.11 Analysis of nuclei velocity

Using the DiGraph of nuclei tracks, the velocity of each nucleus from one time point to the next was calculated. To calculate the velocity of a nucleus, the euclidean distance between the position at t-1 and t was calculated. Then, the distance was divided by the time elapsed in one time point. Finally, the velocities of each nucleus were plotted using SeaBorn³⁷.

5.2.12 Mercator projection of nuclei onto a 2D space

The position of the nuclei in the embryo was projected onto a 2D image. First, given the spherical shape of *P. hawaiiensis* embryos, a sphere was fitted by using a Least Square Sphere Fit³⁸. Then, using this sphere model (center and radius), a Mercator projection was performed³⁹ using only the Phi and Theta parameters of each nucleus (thereby projecting the nuclei onto the sphere and removing their offset to the sphere shell). Finally, using the X and Y coordinates obtained from the projection, I plotted each time point using SeaBorn³⁷.

5.2.13 Rendering of embryos in volumetric space with Blender

I used a procedure that I had previously developed for rendering a volumetric dataset in Blender and made publicly available:

<https://hackmd.io/@sOXXFIraQiiB2hiInQpyMw/rketOixDG?type=view>

5.2.14 Additional scripts created to manipulate file formats used in light sheet microscopy

All additional scripts that I created to perform these analyses are found in the GitHub repository https://github.com/extavourlab/Blondel_Leo_Thesis under scripts/Chapter_4/Scripts/light sheetUtils

- MaMuT.py: A python library to open and manipulate MaMuT XML files. Exposes the parameters and allows for the resaving of XML files.
- MaMuTUtils.py: A set of tools to manipulate MaMuT XML files as a Python library. It can:
 - Merge multiple MaMuT XML files.
 - Append a set of annotations to an existing XML file.
 - Change the path of the microscopy images associated with a set of annotations
 - Change the number of time points in a dataset
 - Clean unconnected annotations (annotations without a track)
 - Remove annotations given a set of parameters (position or size)
 - Merge colocalized spots
- hdf5_Make_BDVXML.py: Creates a BigDataViewer XML file from a BigDataViewer HDF5 file.
- hdf5_to_8RAW.py: Converts a BigDataViewer HDF5 file into an 8-bit RAW file. Allows for the selection of time points, and resolution. Useful for importing into OpenGL volumetric renderer such as Blender.
- hdf5_to_TIFF.py: Converts a BigDataViewer HDF5 file into a TIFF stack. Allows for the selection of time points, and resolution.
- bdvxml.py: Open and parses a BigDataViewer XML file for manipulation. Allows for resaving of manipulated XML files.
- HDF5_Merge_multiple.py: Merges multiple BigDataViewer HDF5 files into one.
- CleanUnlaid.py: Remove Unlaid spots in a MaMuT XML file. Unlaid spots are annotations that do not belong to any track.

- `ExtractPixelsFromMaMuT.py`: Uses the MaMuT annotations to extract small image cubes containing a single nucleus. Useful for subsequent training, or visualization of the quality of the annotations.

5.3 Results

5.3.1 Imaging of nuclei and membrane of wild type *P.*

hawaiensis embryos with the light sheet SimView scope

As discussed in the introduction of Chapter 4, the main goal of this chapter was to generate a geo-localization system for transcripts, on a live developing embryo. To achieve this goal I needed to generate high-quality volumetric recordings of developing embryos and track every cell.

Thanks to Anastasios Pavlopoulos, I was invited to record the early embryogenesis of *P. hawaiensis* at the HHMI Janelia Research Campus, in the laboratory of Philip Keller. The Keller laboratory specializes in light sheet microscopy and has developed multiple cutting edge light sheet microscopes^{7,16,40}. For all of the experiments described in this chapter, I used the SimView⁷ microscope.

Due to the invariant early cleavages which allow for the unambiguous identification of each lineage up until the 16 cell stage, the beginning of the recording was set to be at 8 cells (stage S4), with an error margin up to the 16 cell stage (beginning of stage S5). As mentioned in Chapter 4, the transgenic line *PhHS>H2B-mRFP**ruby* could not be used for the recording of early embryogenesis due to the inability to activate the heat shock element until the S6 stage. Therefore I used transient expression of marker proteins through the micro-injection (performed by E. Stamatakis for the 2017 datasets and by me for the 2018 datasets) of messenger RNA⁴¹. One cell stage (stage S1) embryos were injected with two different mRNAs, one encoding the nuclear protein Histone 2 A (H2A) tagged with the fluorescent reporter mCherry (H2A-mCherry), and the second encoding the membrane localization domain Lyn tagged with the fluorescent reporter GFP (Lyn-GFP) (Figure 5.4). One cell stage embryos injected with these

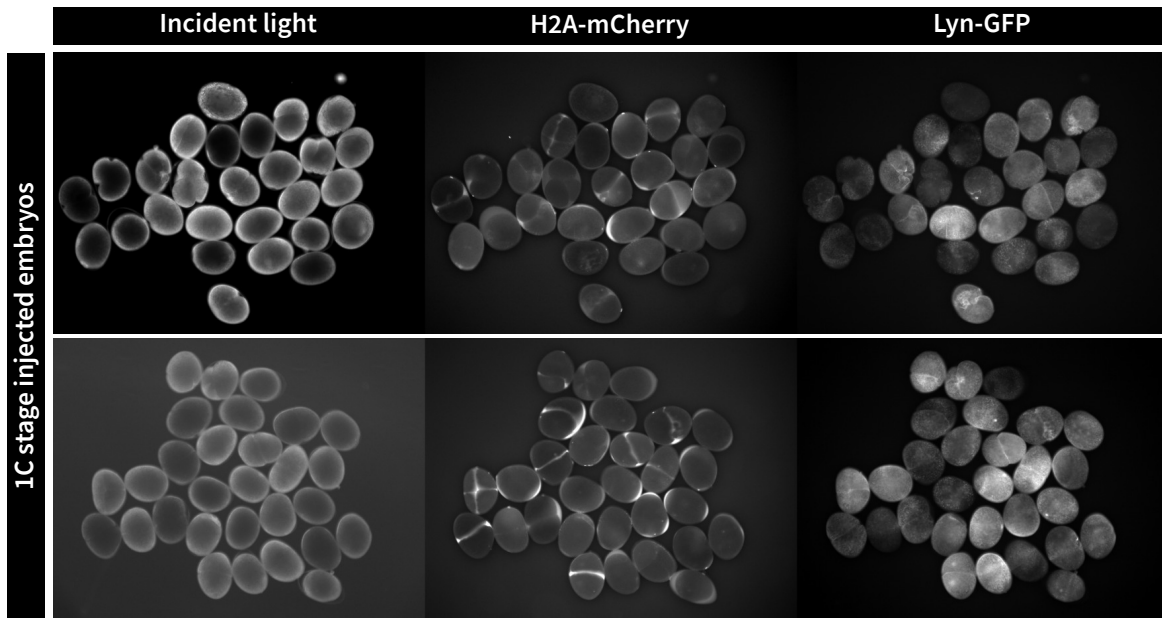


Figure 5.4: Representative images of *P. hawaiensis* injected embryos. Embryos were injected at the one-cell stage with mRNA for *H2A-mCherry* and *Lyn-GFP* and imaged using a stereomicroscope 4h later (when they are at the one, two, or four cell stage). Embryos are showing different expression levels along with stages from the 2 to 4 cell stages.

mRNAs had an 81%(+8%) survival rate. For each of two replicate experiments (Table 5.1), three successfully injected embryos were mounted and imaged on the SimView scope using a custom mounting procedure I developed (see Methods: *Preparation of embryos for light sheet microscopy with the SimView Keller laboratory scope*). For every experiment, a unique set of parameters within the constraints of light sheet microscopy was chosen to maximize the number of angles while reducing light damage and to have a small time-resolution to allow for accurate nuclei tracking (Table 5.1). In total, I successfully recorded four wild type embryos, in two experiments, expressing *H2A-mCherry* and *Lyn-GFP* (Table 5.1). For both experiments, out of the three embryos I mounted, one failed to develop fully and stopped developing during the recording. The images were then processed using the FIJI Multiview Reconstruction pipeline¹⁷. The registration was done successfully on each dataset using the beads embedded in the agar with a pixel error rate of 2.5(+0.5)px. Each dataset was then fused, using the weighted-average fusion algorithm¹⁷, resulting in a 0.406um isometric voxel resolution. Representative images of the fused datasets are shown in (Figure 5.5).

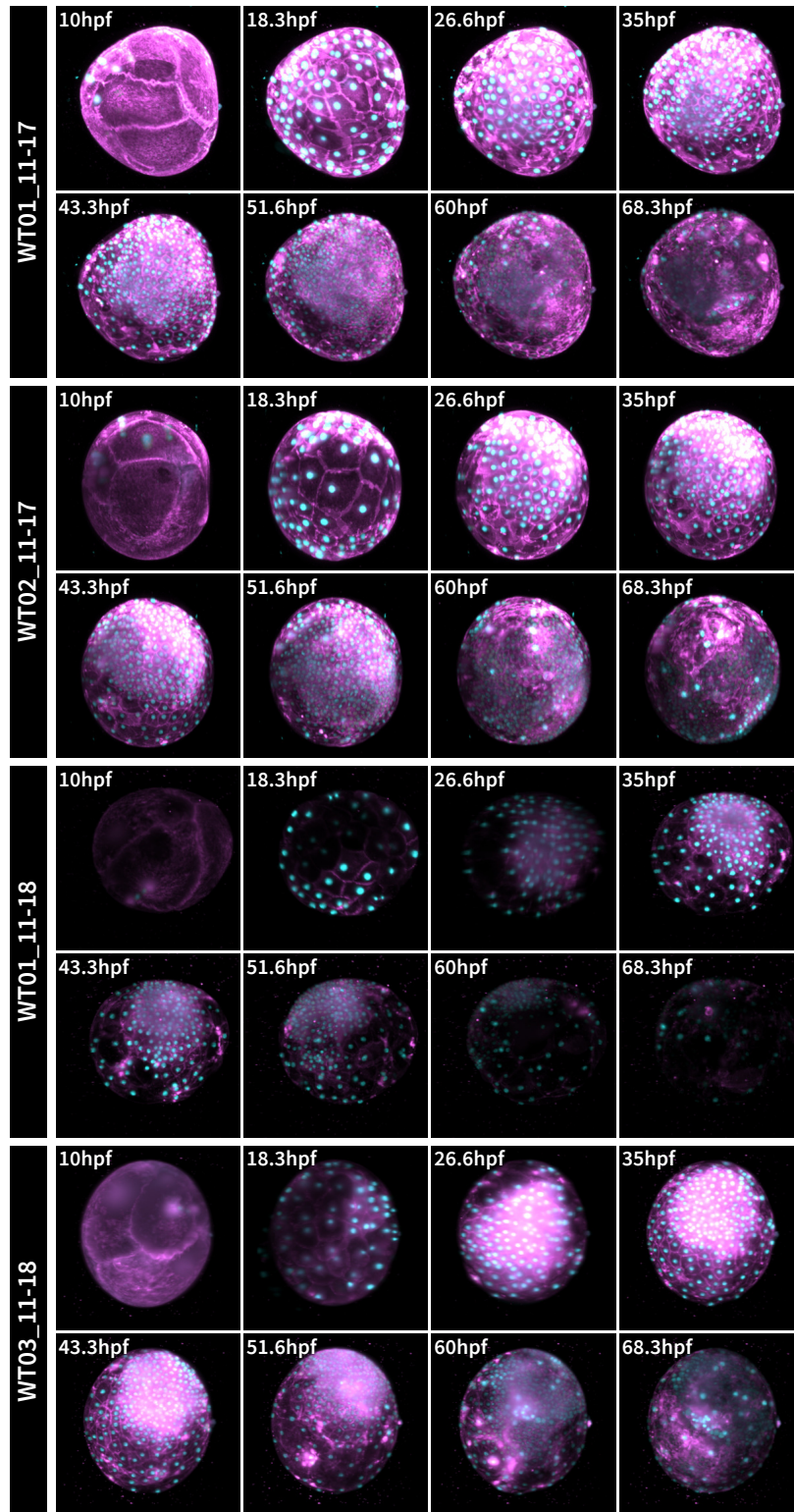


Figure 5.5: Maximum intensity projection of fused light sheet recording of *P. hawaiiensis* embryos. Projections were computed for each embryo from the same angle and every 8.3 hours of recording. WT01_11-18 and WT02_11-18 recordings continue for another 11.5 hours but are not shown here. In cyan are the nuclei tagged with H2A-mCherry and in magenta is the membrane tagged with Lyn-GFP.

Date	Name	Condition	Angles (equally spaced)	time points	Time resolution	Status
11-2017	WT_01-11-17	Wild Type	8 angles	366	10mn	Fused and Tracked
11-2017	WT_02-11-17	Wild Type	8 angles	366	10mn	Fused
11-2017	WT_03-11-17	Wild Type	8 angles	366	10mn	Deleted due to embryo death
11-2018	WT_01-20-11-18	Wild Type	6 angles	419	10mn	Fused
11-2018	WT_02-20-11-18	Wild Type	6 angles	419	10mn	Deleted due to embryo death
11-2018	WT_03-20-11-18	Wild Type	6 angles	419	10mn	Fused

Table 5.1: List of every wild type *P. hawaiiensis* embryo recorded with the SimView microscope. Important dataset metadata are written in the corresponding columns. The status of the processing is reported in the "Status" column. Images of embryos shown here are visible in Figure 5.5.

5.3.2 Analysis of nuclei movement and division dynamic

To track the exact position of the nuclei in space the first step I took was the segmentation of the nuclei in the embryos. Multiple methodologies for the segmentation of nuclei in 2D have been developed over time^{25,42,43,44,45}. Using Ilastik²², I trained a random forest model with the training interface to segment the nuclei. Due to the changes in fluorescence intensity, three different models had to be trained, one for the first 100 time points, one for the next 100, and one for the last 100. Due to the high amount of time required to segment nuclei with Ilastik, I trained, tested, and segmented only the first dataset (WT_01-11-17) to assess the viability of this algorithm. Finally, I used the segmented model as an input to the object tracking pipeline of Ilastik and exported the results to visually assess their quality using the MaMuT plugin¹¹.

The results contained many erroneous annotations. First, a large number of objects were unconnected and corresponded to fluorescent artifacts in the images (Figure 5.6). Second, the tracks showed a heavy level of fragmentation, with over 1062 generated tracks for 16 originating cells (Figure 5.6). I first wrote a small algorithm that would remove any annotation that was not

part of a track (unconnected annotations) to reduce the amount of noise. Then I manually corrected all of the tracks from the first time point to time point 300. While performing this task, I also cleaned any tracks that followed fluorescent artifacts, along with tracks tracking fluorescent beads. This resulted in a clean set of annotations for all nuclei in the WT01-11-17 dataset (Figure 5.6). The remaining embryos were not segmented or tracked. However, WT01_11-17 is a fully tracked positive control that can be used as a comparison to any new automated segmentation and tracking algorithm.

Using the fully tracked WT_01-11-17 embryo I first asked if there was any pattern to the nuclei's division dynamics across the different stages of embryogenesis recorded, namely early division, gastrulation, germ disk aggregation, and germ band formation. I extracted the number of nuclei in each frame and computed the rolling hourly average. Due to the exponential nature of cellular divisions, I plotted the Log₂ of the number of nuclei against time. Interestingly, I observed four different rates of division corresponding to four different stages of development (Figure 5.7). In the first half of stage S5, from 16 to 64 cells (Early divisions), the rate of division is the highest. It then slows in the second half of stage S5 between 64 and 128 cells (Cell migration) (Figure 5.7). During gastrulation, cellular division seemingly stops (Figure 5.7). Finally, the rate of division resumes at a slower pace during the formation of the germ band (Figure 5.7). Moreover, for the first three different division phases, the nucleus velocities change concurrently. During the early divisions, the velocity of nuclei slowly drops. I speculate that this may be due to the fact that the volume occupied by each cell halves at each division, such that the total distance traveled halves twice, leading to a reduction in speed from 90um/h to 30um/h. During the next phase, the velocity increases. This phase is called the yolk segregation phase³¹ and corresponds to a reduction in the size of the cells due to the extrusion of the yolk from the forming cells, as well as the start of active cellular movement^{31,35,46}. Finally, during gastrulation, the motion of cells diminishes to what will become the baseline cellular velocity during germ band formation. While these results are robust within this embryo, they must be subjected to replication in future studies, since only a single embryo was used to assess these dynamics.

I then asked what the contribution of each germ layer lineage was to the total amount of cells in the embryo during early embryogenesis. I first plotted each track as a tree, where all lines

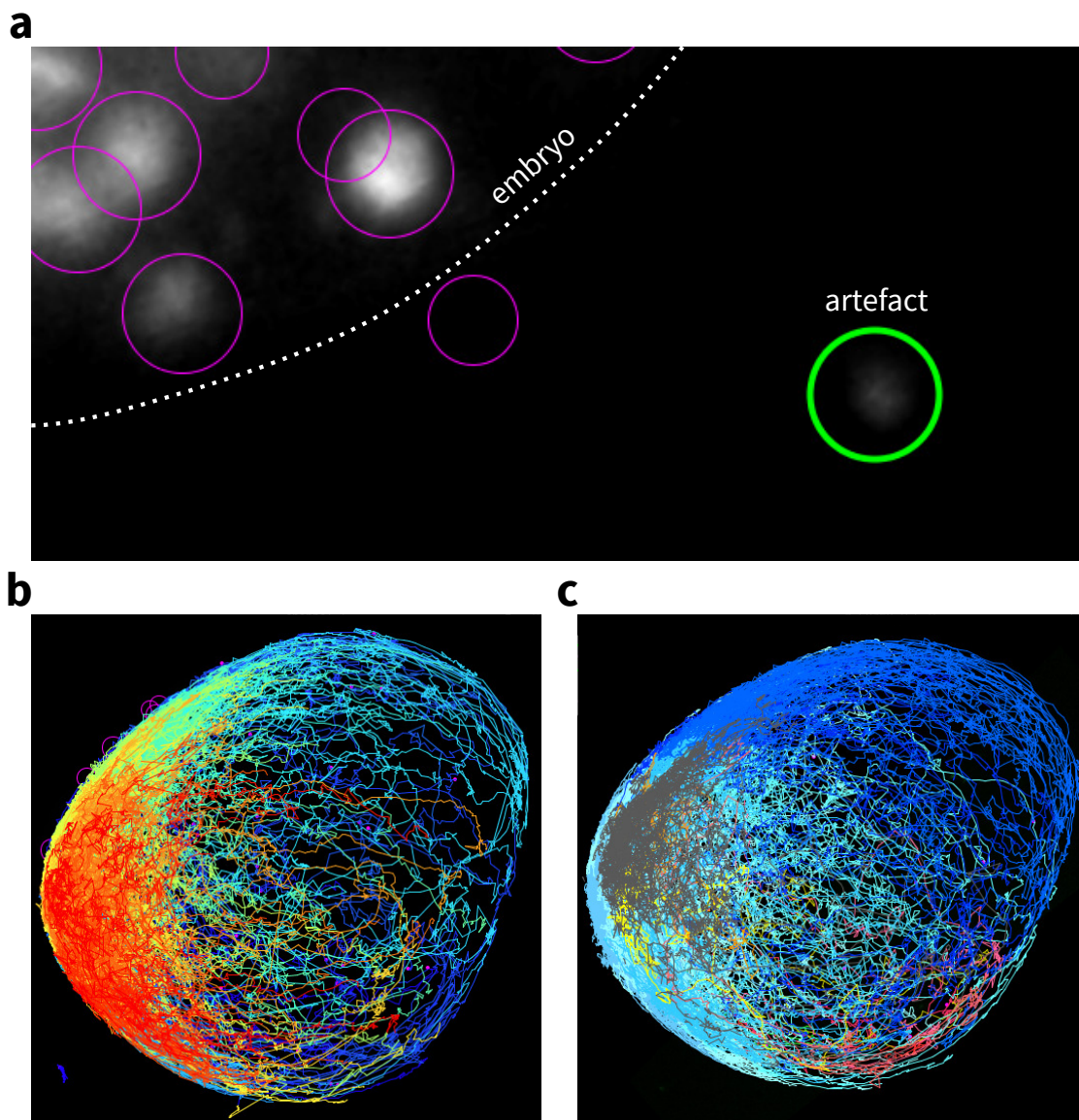


Figure 5.6: Results from the automated segmentation and tracking performed using Ilastik and manual tracking and corrections. **a)** Example of artefactual segmentation by the automated process. Each circle represents an annotation, the green circle shows the artifact. **b)** Tracks generated by the flow tracking algorithm of Ilastik. Each unique track is given a color from blue to red. 1062 unique tracks are represented on this embryo. **c)** Tracks remaining after manually correcting the tracks from (b) using MaMuT. Each track is colored by its blastoderm origin: blues are El Er and Ep, yellow is g, green is en, reds are ml, and mr and orange is Mav. In gray are tracks for which the blastoderm origin is unknown due to the nuclei becoming invisible while traveling inside the yolk.

represent a cell and are colored by the originating blastomere (Figure 5.8). I then calculated the contribution of each blastomere to the cell population at 60hpf, as well as the contribution of each germ layer (Figure 5.8). The ectoderm contributed the majority (79%) of cells. This was expected, as most of the mesoderm is generated during the segmentation of the germ band, which only starts towards the end of the recording^{32,35,47}. I then plotted the contribution of each blastomere lineage and germ layer to the cell population over time (Figure 5.8). At first, the mesoderm and ectoderm lineages contribute equally to the cell population (by the 8 and 16 cell stages, three and six (respectively) of the blastomeres are from the mesoderm and ectoderm lineages^{32,35}) (Figure 5.8). However, by the time gastrulation starts, the proportion of cells is already heavily skewed towards the ectodermal lineage, with 77% of the cells coming from El, Er, or Ep (Figure 5.8). This proportion maintains itself until 52hpf, when the contribution from other lineages starts to increase again (Figure 5.8). Importantly, no descendants of El, Er, and Ep disappeared, and all were fully tracked. Therefore the contribution from unknown lineages cannot come from the ectoderm. However, due to the loss of cells to the tracking process from the other lineages during gastrulation, I was not able to assign a direct lineage identity to 19% of the population by 60hpf (Figure 5.8).

Finally, I characterized two major cellular rearrangements during gastrulation and the formation of the germ band. First, right before gastrulation, the cells go from a state of random motion to one of directed motion towards the anterior pole of the embryo (Figure 5.9). This pole is located at the point where the descendants of El, Er, and Ep cross (Figure 5.9). This directed motion is followed by the ingression of the descendants of Mav and en between cells coming from El and Er. Mav and en descendants subsequently migrate inside the embryo forming a second layer of cells under the descendants of El and Er, which migrate over that area via epiboly (Figure 5.9). Second, the cells from Ep that will form the midline lying between El and Er descendants, comes from a small group of precursor cells. These precursor cells are slowly engulfed between descendants of El and Er. However, the lineage boundaries are already present after gastrulation and slowly refine themselves subsequently (Figure 5.9).

Without the study of more embryos to confirm those results, it is not possible to conclude whether the observed cell rearrangement, cellular territories, lineages proportion, division

rates, and nuclei dynamic observed here are variable or conserved between embryos. However, this pilot experiment demonstrates the feasibility of these methods to successfully image, track, and analyze the early development of *P. hawaiiensis* at single-cell resolution.

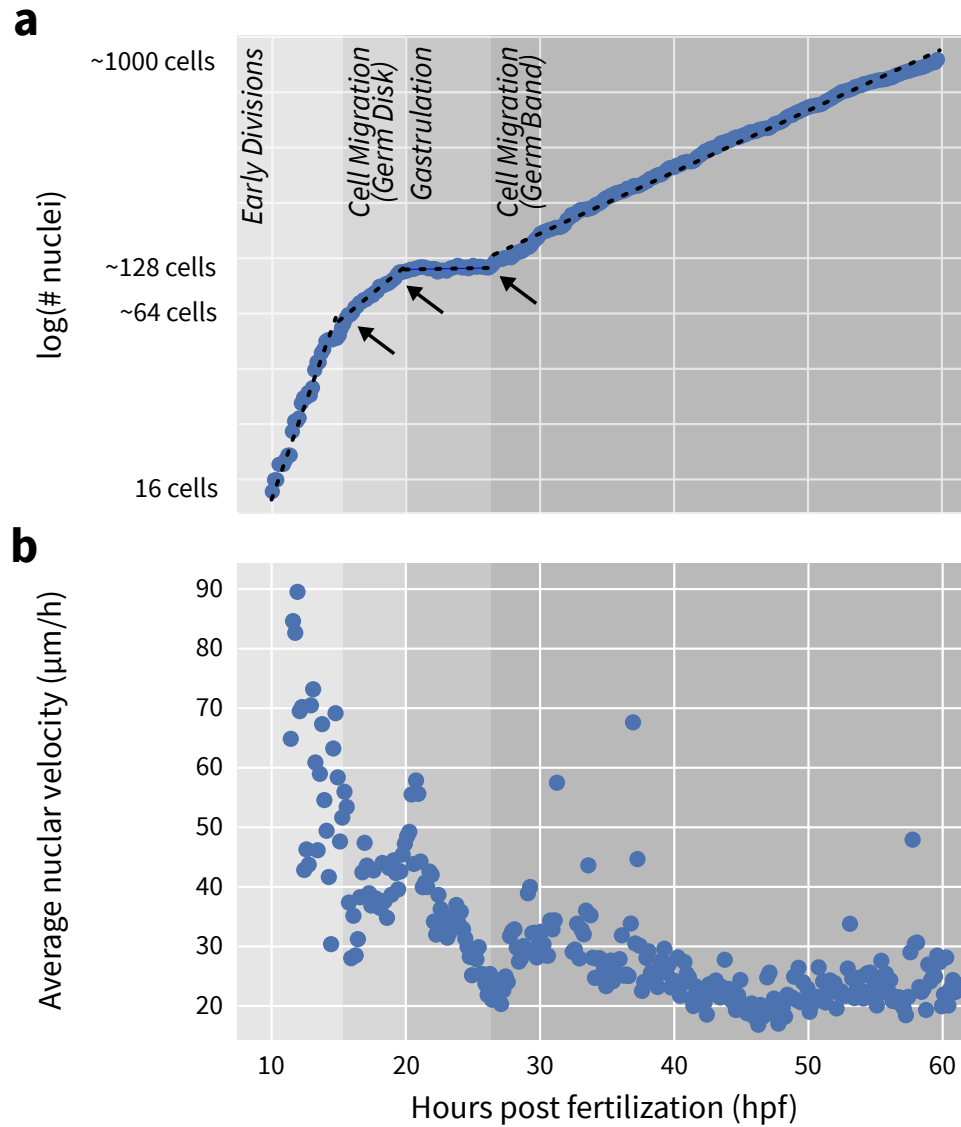


Figure 5.7: Analysis of cell division rates and nuclear velocities in the tracked WT01_11-17. Four different key developmental processes are represented by grey backgrounds. Each process is annotated in **a**. **a**) 1h rolling average of the number of nuclei at each time point. Due to the exponential nature of cell division, the number of nuclei is represented on a log scale. Black dotted lines represent the linear regression for the number of nuclei. Each regression was performed on the subset of time points representing a developmental process. All four regressions have a p-value < 0.05. Inflection points are shown by black arrows. **b**) Average velocity of nuclei at each measured time point. The developmental stages are shown in grey following the order and annotation drawn in **a**.

Figure 5.8 (following page): Analysis of blastoderm and germ lineage dynamics during the early embryogenesis of *P. hawaiiensis*. **a)** Tree representation of blastoderm lineages derived from the WT01_11-17 tracked embryo. Each blastomere is colored according to the legend at the bottom of the figure. Each division generates a horizontal line while time goes from top to bottom. If a track stops it means that the cell was lost during tracking. All tracks that could not be mapped back to a blastomere were removed from this representation. **b)** Analysis of the number of cells coming from each blastoderm lineage at 60hpf. The total number of cells was counted and plotted on a bar plot. Cells from unknown lineages (ukn) are displayed in grey. **c)** Analysis of the number of cells participating in each germ layer at 60hpf. The total number of cells was counted and plotted on a bar plot. **d)** Contribution of each blastoderm lineage to the total number of embryonic cells during the first 60 hours of embryogenesis. At each time point, the number of cells from each lineage was extracted from the annotations and the ratio to the total number of cells in the embryo computed. Each ratio was then plotted on a time stackplot and each area was colored by the blastoderm lineage. **e)** Contribution of each germ layer to the total number of cells during the first 60 hours of embryogenesis. At each time point, the number of cells from each germ layer was extracted from the annotations and the ratio to the total number of cells in the embryo computed. Each ratio was then plotted on a time stackplot and each area was colored by the germ layer identity.

Figure 5.8: (continued)

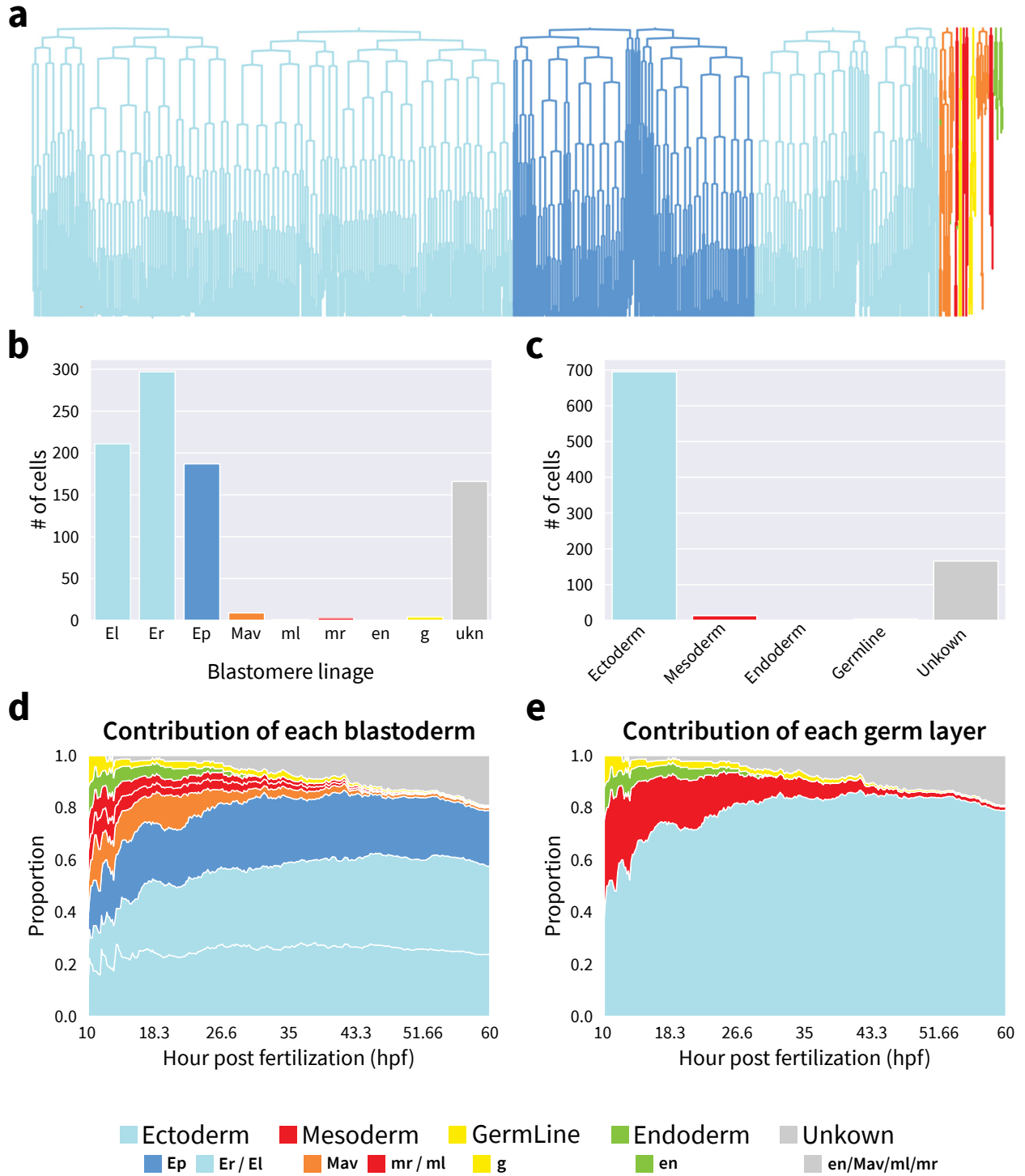
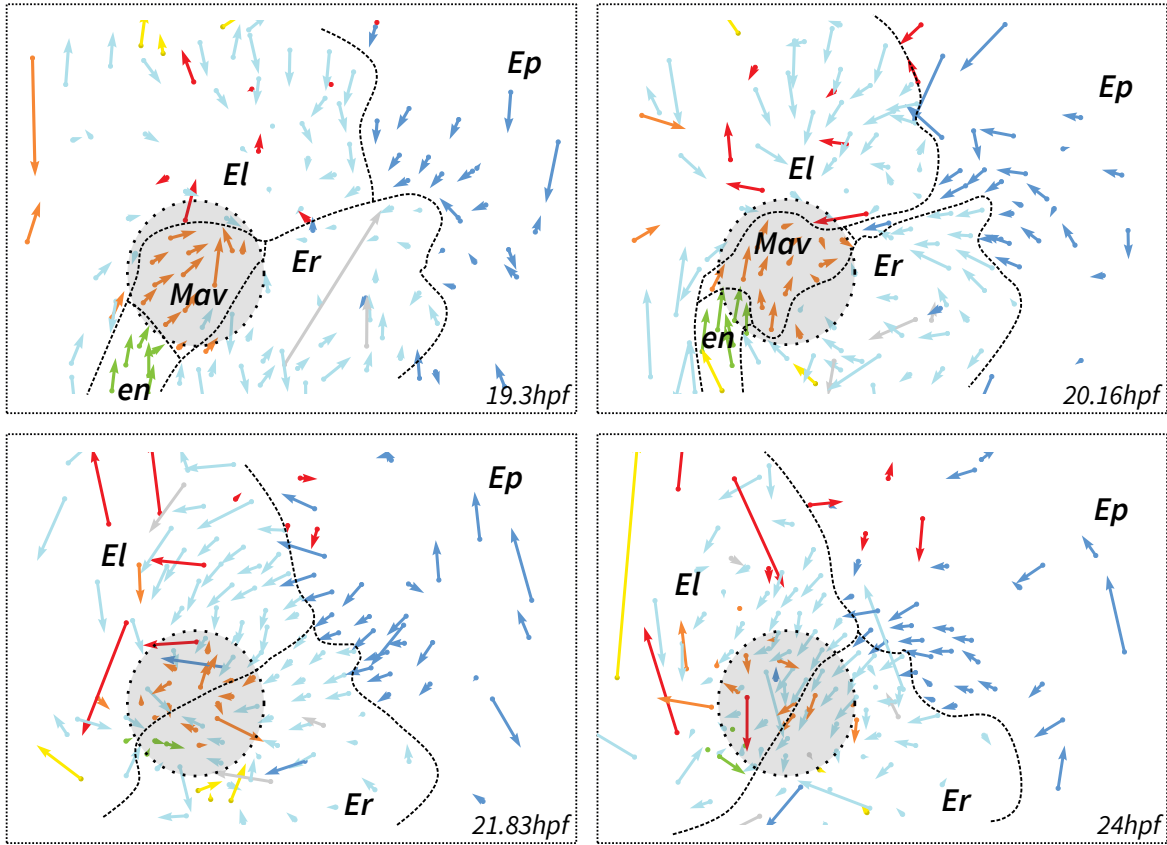


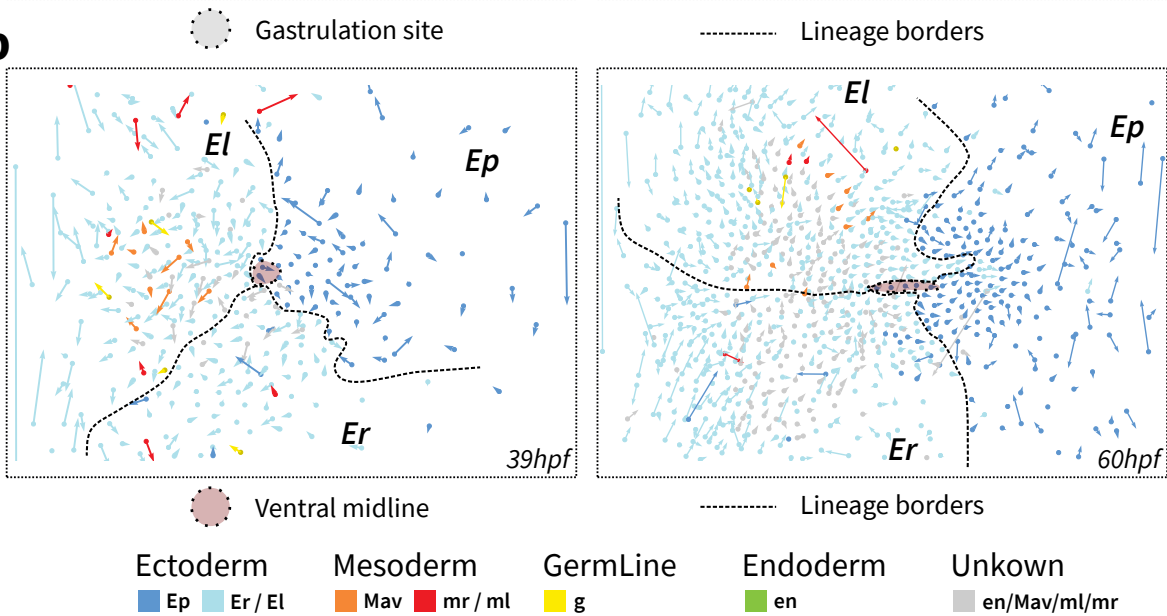
Figure 5.9 (following page): Analysis of cellular movement during gastrulation and formation of the midline. A sphere was mapped onto the embryo and used to project each cell position to a 2D representation via a Mercator projection. Cells are colored by their blastoderm lineage. Blastoderm lineage borders of cell territories are schematized using dashed lines and text annotation. The site of gastrulation is represented by a greyed dashed circle. The ventral midline is represented by a red shade and dashed border. Cell movement is represented by a vector; the direction of the vector is the direction of movement and the length of the vector is proportional to the velocity of the cell between this time point and the next. **a)** Visualisation of gastrulation on the Mercator projections. Four representative time points of gastrulation were selected and annotated here. **b)** Visualization of the formation of the ventral midline. Two representative time points of midline migration and formation were selected and annotated here.

Figure 5.9: (continued)

a



b



5.3.3 The effect of loss of a blastomere on morphogenesis

As mentioned before, upon ablation of one of the eight cell stage blastomeres from the ectodermal or mesodermal lineages, the embryo is capable of compensating for the missing lineage by the germ band stage (Chapter 4: Figures 4.2 and 4.3)⁴⁸. This display of intra-germ layer compensation may seem at first counterintuitive given the high level of stereotypic divisions in the early embryo^{32,35,48} (Figure 5.10). To better understand how such compensation could occur, I designed in collaboration with Anastasios Pavlopoulos and Cassandra Extavour an experiment to study the compensation of the loss of one of the right or left Ectoderm lineage founder blastomeres (El or Er). We chose to ablate El or Er for two main reasons: First, as we have seen above, the ectoderm lineage remains at the surface of the embryo and is, therefore, easier to track and image with light sheet microscopy. Second, the Er (or El) cells are regenerated from both the Ep and El (or Er) cell lineages (⁴⁸). This implies that two embryonic body plan boundaries must be crossed, the Anterior-Posterior boundary with cells from the Ep lineage invading the anterior part, and the Right-Left boundary with cells from El (resp. Er) invading the right side of the embryo or vice versa. Using the previously developed blastomere ablation technique³³ Cassandra Extavour and I ablated the Er blastomere in embryos injected with tracer mRNA. The first part of this experiment was done exactly as described in the previous section. Then, at the eight cell stage, we performed the ablation of Er, then waited for one hour for the embryo to recover (Figure 5.10) before mounting and imaging three embryos per experiment on the SimView scope as described previously. In total six successful ablations were recorded in two experiments (Table 5.2). The datasets were then registered successfully. One of the embryos was lost during the recording at time point 124. To date, no further analysis has been performed on those datasets.

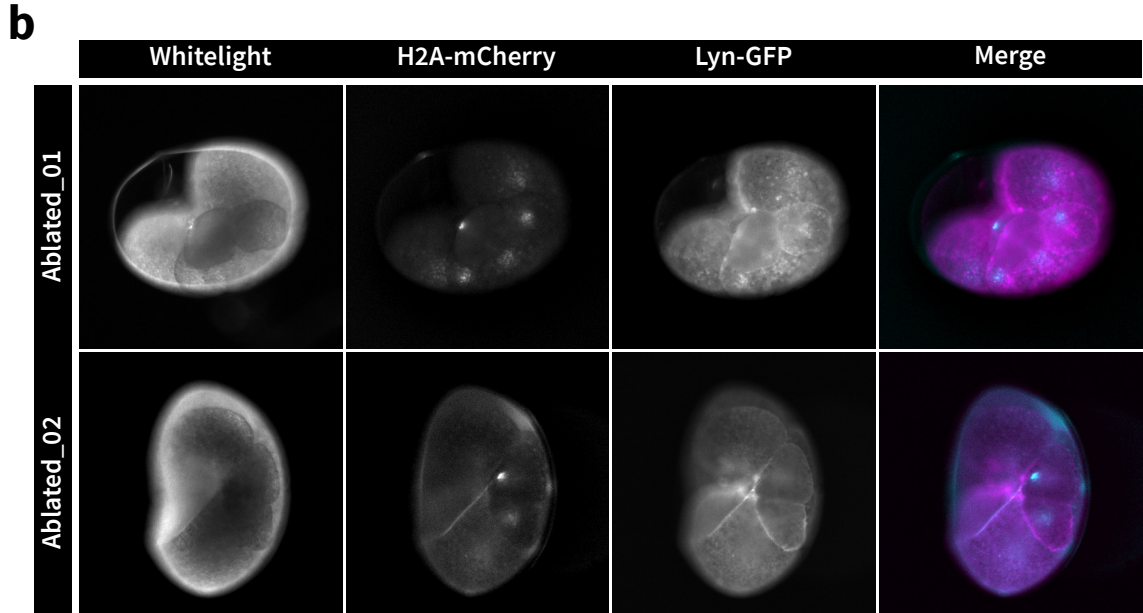
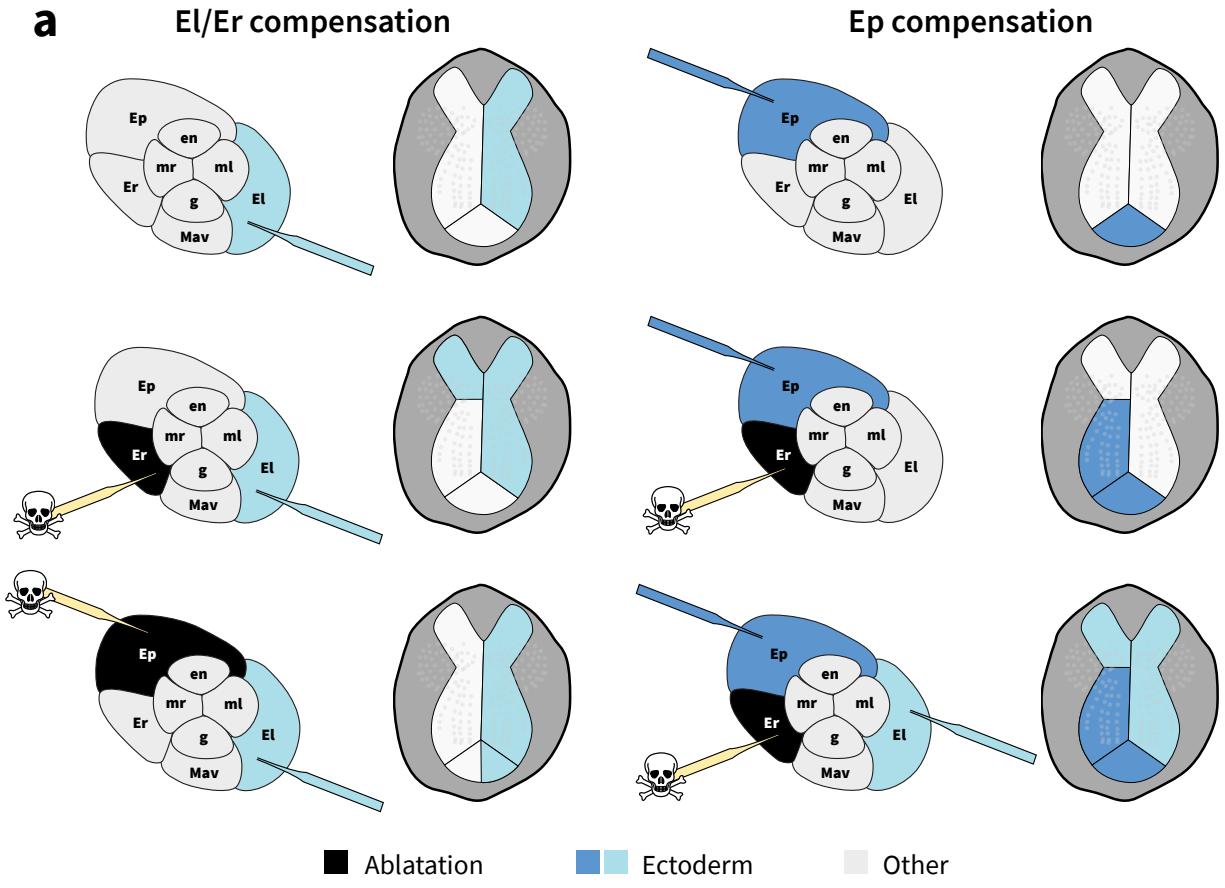


Figure 5.10: Schematic and visual representation of the blastomere ablation experiments. **a)** Schematic representation of the ablation experiment adapted from⁴⁸. One of the three ectodermal blastoderms, El, Er, and Ep, was ablated while another blastomere was injected with a dye. The results of the experiments are presented in the schematic representation of germband embryos. **b)** Representative images of the blastomere ablation experiments performed at the Janelia Research Campus. Two of the 6 recorded embryos are shown here, showing an ablation of Er. Embryos were injected at the one-cell stage with mRNA for H2A-mCherry and Lyn-GFP.

Date	Name	Ablated cell	Angles (equally spaced)	time points	Time resolution	Status
11-2018	Ablated_01	Er	6 angles	433	10mn	Registered
11-2018	Ablated_02	Er	6 angles	433	10mn	Registered
11-2018	Ablated_03	Er	6 angles	330	10mn	Registered
11-2018	Ablated_04	Er	6 angles	124	12mn	Registered
11-2018	Ablated_05	Mav	6 angles	331	12mn	Registered
11-2018	Ablated_06	Er	6 angles	331	12mn	Registered

Table 5.2: List of all embryos successfully ablated and imaged using the SimView microscope. The date and name of each embryo are reported along with the identity of the ablated blastomere. Metadata regarding the dataset are reported in the subsequent columns. Finally, the current status of the dataset is reported in the last column.

5.4 Discussion

In this chapter, I described the recording of the early embryogenesis of *P. hawaiiensis* using light sheet microscopy. This project generated high-quality datasets of wild type embryos developing for three or four days. However, I did not finish the complete tracking and analysis of all embryos. As described in the introduction, the challenges of automating the analysis of three-dimensional datasets are not yet solved. Despite this, I generated full tracks of a wild type embryo that can now be used to try and test any algorithm against. I believe that automated approaches will be needed in the future due to the extensive amount of time necessary to track a single embryo from start to finish. For any comparative study involving a sufficient amount of embryos for statistics to be derived, I believe that a fully (or almost fully) automated pipeline must be developed. The embryo that I tracked can be used to train a deep learning model thanks to the vast amount of data contained in a single dataset. For example, considering only the last 100 time points of the dataset, the 1000 nuclei per frame, the three dimensions of the cube, and image augmentation, it would be possible to generate one million annotated nuclei to train a deep learning algorithm. New methods such as the star-convex deep learning algorithm StarDist²⁵ could in principle be trained with 1% of the potential data annotated here, given that successful segmentation is reported with as few as four training volumes containing >1000 nuclei²⁵).

Beyond the segmentation of nuclei, the tracking of nuclei in three dimensions is a current challenge²⁰. New object tracking algorithms in three dimensions have been proposed, such as

the time-reversed flow algorithm used in the tracking of the mouse embryo in the Keller laboratory¹⁰, or the Baxter algorithm⁴⁹. Using tracks I generated for my embryonic dataset WT01_11-17, it should be possible to assess the accuracy of those algorithms and to tune their hyperparameters such that optimal tracking can be achieved.

In the final part of this chapter, I discussed the advances that I made towards understanding the processes underlying the intra-germ layer regeneration. By using the same imaging techniques that were successful for the wild type embryos, and by leveraging the ablation protocol developed previously in the lab, I generated a large microscopy dataset of regenerating embryos. The tracking of nuclei in these datasets will need to be performed in the future, and should be automatized as discussed above. This dataset holds the potential for fascinating discoveries on morphogenesis and the establishment of cellular territories. By comparing the motions and positions of cells of the ectodermal lineage, hypotheses towards the establishment of the strict boundaries may be established. Similarly, future studies of what other processes may be involved in the compensatory mechanisms will shine a light on this regenerative property.

I note that I used only half of the data generated, namely the nuclear information, and did not analyze the membrane channel. Another current Ph.D. student in the Extavour laboratory, Beatrice Steinert, has successfully used the surface projection tool IMSane⁵⁰ to create two-dimensional projections of the membranes of embryonic dataset WT01_11-17. This information can be used in the future to study the morphogenetic changes happening during the development of *P. hawaiiensis*. For example, these data can help to determine the cellular rearrangements that are required for the establishment of the grid stage.

Finally, the experiments of both the preceding chapter and this chapter have been the most challenging I have ever executed. The diversity of skills required for the successful completion of this project was probably, in hindsight, too high for a single person. Moreover, the amount of time required for each step made this project complex to finish within the timespan of a Ph.D. Nevertheless, this has also been extremely inspiring, and I consider the successes regarding the collection of rich light sheet datasets in itself a great contribution. I look forward to the completion of the analysis of those datasets by future researchers.

References

- [1] H. Siedentopf and R. Zsigmondy. Über Sichtbarmachung und Größenbestimmung ultramikroskopischer Teilchen, mit besonderer Anwendung auf Goldrubingläser. *Annalen der Physik*, 315(1):1–39, 1902.
- [2] J. W. Strutt. LVIII. On the scattering of light by small particles. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science.*, 41(275):447–454, 1871.
- [3] J. Li, Y. Yin, S. Fortunato, and D. Wang. A dataset of publication records for Nobel laureates. *Sci Data*, 6(1):33, April 2019.
- [4] P. J. Keller and H.-U. Dodt. Light sheet microscopy of living or cleared specimens. *Curr. Opin. Neurobiol.*, 22(1):138–143, February 2012.
- [5] C. J. R. Sheppard. Advanced Light Microscopy. Vol. 2. Specialized Methods. *J. Mod. Opt.*, 37(7):1277–1278, July 1990.
- [6] J. Huisken, J. Swoger, F. Del Bene, J. Wittbrodt, and E. H. K. Stelzer. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science*, 305(5686):1007–1009, August 2004.
- [7] R. Tomer, K. Khairy, F. Amat, and P. J. Keller. Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy. *Nat. Methods*, 9(7):755–763, June 2012.
- [8] P. J. Keller, F. Pampaloni, and E. H. Stelzer. Life sciences require the third dimension. *Curr. Opin. Cell Biol.*, 18(1):117–124, February 2006.

- [9] P. J. Keller, A. D. Schmidt, A. Santella, K. Khairy, Z. Bao, J. Wittbrodt, and E. H. K. Stelzer. Fast, high-contrast imaging of animal development with scanned light sheet-based structured-illumination microscopy. *Nat. Methods*, 7(8):637–642, August 2010.
- [10] K. McDole, L. Guignard, F. Amat, A. Berger, G. Malandain, L. A. Royer, S. C. Turaga, K. Branson, and P. J. Keller. In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. *Cell*, 175(3):859–876.e33, October 2018.
- [11] C. Wolff, J.-Y. Tinevez, T. Pietzsch, E. Stamatakis, B. Harich, L. Guignard, S. Preibisch, S. Shorte, P. J. Keller, P. Tomancak, and A. Pavlopoulos. Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. *Elife*, 7:e34410, March 2018.
- [12] F. Strobl, S. Klees, and E. H. K. Stelzer. Light Sheet-based Fluorescence Microscopy of Living or Fixed and Stained *Tribolium castaneum* Embryos. *J. Vis. Exp.*, (122):55629, April 2017.
- [13] A. Kaufmann, M. Mickoleit, M. Weber, and J. Huiskens. Multilayer mounting enables long-term imaging of zebrafish development in a light sheet microscope. *Development*, 139(17):3242–3247, September 2012.
- [14] T. Ichikawa, K. Nakazato, P. J. Keller, H. Kajiura-Kobayashi, E. H. K. Stelzer, A. Mochizuki, and S. Nonaka. Live imaging and quantitative analysis of gastrulation in mouse embryos using light-sheet microscopy and 3D tracking tools. *Nat. Protoc.*, 9(3):575–585, March 2014.
- [15] W. C. Lemon, S. R. Pulver, B. Höckendorf, K. McDole, K. Branson, J. Freeman, and P. J. Keller. Whole-central nervous system functional imaging in larval *Drosophila*. *Nature Communications*, 6(1), 2015.
- [16] R. K. Chhetri, F. Amat, Y. Wan, B. Höckendorf, W. C. Lemon, and P. J. Keller. Whole-animal functional and developmental imaging with isotropic spatial resolution. *Nat. Methods*, 12(12):1171–1178, December 2015.
- [17] S. Preibisch, S. Saalfeld, J. Schindelin, and P. Tomancak. Software for bead-based registration of selective plane illumination microscopy data. *Nat. Methods*, 7(6):418–419,

June 2010.

- [18] S. Preibisch, S. Saalfeld, and P. Tomancak. Globally optimal stitching of tiled 3D microscopic image acquisitions. *Bioinformatics*, 25(11):1463–1465, June 2009.
- [19] S. Preibisch, F. Amat, E. Stamatakis, M. Sarov, R. H. Singer, E. Myers, and P. Tomancak. Efficient Bayesian-based multiview deconvolution. *Nat. Methods*, 11(6):645–648, June 2014.
- [20] V. Ulman, M. Maška, K. E. G. Magnusson, O. Ronneberger, C. Haubold, N. Harder, P. Matula, P. Matula, D. Svoboda, M. Radojevic, I. Smal, K. Rohr, J. Jaldén, H. M. Blau, O. Dzyubachyk, B. Lelieveldt, P. Xiao, Y. Li, S.-Y. Cho, A. C. Dufour, J.-C. Olivo-Marin, C. C. Reyes-Aldasoro, J. A. Solis-Lemus, R. Bensch, T. Brox, J. Stegmaier, R. Mikut, S. Wolf, F. A. Hamprecht, T. Esteves, P. Quelhas, Ö. Demirel, L. Malmström, F. Jug, P. Tomancak, E. Meijering, A. Muñoz-Barrutia, M. Kozubek, and C. Ortiz-de Solorzano. An objective comparison of cell-tracking algorithms. *Nat. Methods*, 14(12):1141–1152, December 2017.
- [21] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, 7(10):R100, October 2006.
- [22] S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, and A. Kreshuk. ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods*, 16(12):1226–1232, December 2019.
- [23] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng. H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes. *IEEE Trans. Med. Imaging*, 37(12):2663–2674, December 2018.
- [24] F. Amat, W. Lemon, D. P. Mossing, K. McDole, Y. Wan, K. Branson, E. W. Myers, and P. J. Keller. Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nat. Methods*, 11(9):951–958, September 2014.

- [25] M. Weigert, U. Schmidt, R. Haase, K. Sugawara, and G. Myers. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3666–3673, 2020.
- [26] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015.
- [27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [28] C. Stringer, M. Michaelos, and M. Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *bioRxiv*, page 931238, February 2020.
- [29] D. Kao, A. G. Lai, E. Stamataki, S. Rosic, N. Konstantinides, E. Jarvis, A. Di Donfrancesco, N. Pouchkina-Stancheva, M. Sémon, M. Grillo, H. Bruce, S. Kumar, I. Siwanowicz, A. Le, A. Lemire, M. B. Eisen, C. Extavour, W. E. Browne, C. Wolff, M. Averof, N. H. Patel, P. Sarkies, A. Pavlopoulos, and A. Aboobaker. The genome of the crustacean *Parhyale hawaiiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. *Elife*, 5:e20062, November 2016.
- [30] T. Pietzsch, S. Saalfeld, S. Preibisch, and P. Tomancak. BigDataViewer: visualization and processing for large image data sets. *Nat. Methods*, 12(6):481–483, June 2015.
- [31] W. E. Browne, A. L. Price, M. Gerberding, and N. H. Patel. Stages of embryonic development in the amphipod crustacean, *Parhyale hawaiiensis*. *Genesis*, 42(3):124–149, July 2005.
- [32] M. Gerberding, W. E. Browne, and N. H. Patel. Cell lineage analysis of the amphipod crustacean *Parhyale hawaiiensis* reveals an early restriction of cell fates. *Development*, 129(24):5789–5801, December 2002.
- [33] A. R. Nast and C. G. Extavour. Ablation of a single cell from eight-cell embryos of the amphipod crustacean *Parhyale hawaiiensis*. *J. Vis. Exp.*, (85):e51073, March 2014.

- [34] C. G. Extavour. The fate of isolated blastomeres with respect to germ cell formation in the amphipod crustacean *Parhyale hawaiiensis*. *Dev. Biol.*, 277(2):387–402, January 2005.
- [35] F. Alwes, B. Hinchin, and C. G. Extavour. Patterns of cell lineage, movement, and migration from germ layer specification to gastrulation in the amphipod crustacean *Parhyale hawaiiensis*. *Dev. Biol.*, 359(1):110–123, November 2011.
- [36] NetworkX. <https://networkx.github.io/>, . Accessed: 2020-9-22.
- [37] seaborn. <https://seaborn.pydata.org/>, . Accessed: 2020-9-22.
- [38] C. F. Jekel. *Obtaining non-linear orthotropic material models for PVC-coated polyester via inverse bubble inflation*. PhD thesis, Stellenbosch: Stellenbosch University, 2016.
- [39] G. Mercator. *Nova Et Aucta Orbis Terrae Descriptio Ad Usum Navigantium Emendate Accommodata*. 1569.
- [40] M. B. Ahrens, M. B. Orger, D. N. Robson, J. M. Li, and P. J. Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods*, 10(5):413–420, May 2013.
- [41] E. J. Rehm, R. L. Hannibal, R. C. Chaw, M. A. Vargas-Vila, and N. H. Patel. Injection of *Parhyale hawaiiensis* blastomeres with fluorescently labeled tracers. *Cold Spring Harb. Protoc.*, 2009(1):db.prot5128, January 2009.
- [42] L. Kametsky, T. R. Jones, A. Fraser, M.-A. Bray, D. J. Logan, K. L. Madden, V. Ljosa, C. Rueden, K. W. Eliceiri, and A. E. Carpenter. Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. *Bioinformatics*, 27(8):1179–1180, 2011.
- [43] K. Y. Win, S. Choomchuay, K. Hamamoto, and M. Raveesunthornkiat. Comparative Study on Automated Cell Nuclei Segmentation Methods for Cytology Pleural Effusion Images. *J. Healthc. Eng.*, 2018:9240389, September 2018.
- [44] F. Long. Microscopy cell nuclei segmentation with enhanced U-Net. *BMC Bioinformatics*, 21(1):8, January 2020.

- [45] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers. Cell Detection with Star-Convex Polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 265–273. Springer International Publishing, 2018.
- [46] R. C. Chaw and N. H. Patel. Independent migration of cell populations in the early gastrulation of the amphipod crustacean *Parhyale hawaiiensis*. *Dev. Biol.*, 371(1):94–109, November 2012.
- [47] R. L. Hannibal, A. L. Price, R. J. Parchem, and N. H. Patel. Analysis of *snail* genes in the crustacean *Parhyale hawaiiensis*: insight into *snail* gene family evolution. *Dev. Genes Evol.*, 222(3):139–151, May 2012.
- [48] A. L. Price, M. S. Modrell, R. L. Hannibal, and N. H. Patel. Mesoderm and ectoderm lineages in the crustacean *Parhyale hawaiiensis* display intra-germ layer compensation. *Developmental Biology*, 341(1):256–266, 2010.
- [49] K. E. G. Magnusson, J. Jalden, P. M. Gilbert, and H. M. Blau. Global linking of cell tracks using the Viterbi algorithm. *IEEE Trans. Med. Imaging*, 34(4):911–929, April 2015.
- [50] I. Heemskerk and S. J. Streichan. Tissue cartography: compressing bio-image data by dimensional reduction. *Nat. Methods*, 12(12):1139–1142, December 2015.

Cyberspace. A consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts... A graphic representation of data abstracted from banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding...

William Gibson, *Neuromancer*, 1989

6

Diving into the third dimension, virtual reality for the analysis of volumetric microscopy

ABSTRACT

From hand drawing to digital images, the way we record microscopy observation has evolved drastically over the last centuries. Light sheet microscopy provides researchers with isomorphic three-dimensional images of biological samples. To observe the datasets generated in Chapter 5, I developed a software that allows the visualization of 3D images in Virtual Reality. By using a Ray Marching rendering shader, volumetric images can be imported and manipulated in a virtual environment. The environment implemented image processing such as contrast and intensity adjustments, slicing, and multi-channel rendering. I designed a virtual interface that allowed the user to change image settings. Moreover, I added the capacity for a user to play through a 3D + time dataset. Finally, I created a visualizer for nuclei tracking annotations to observe the tracks generated in Chapter 5.

ACKNOWLEDGEMENT

I want to thank the members of the Extavour laboratory who tried and tested different iterations of this tool, especially Taro Nakamura who provided numerous datasets and extensively explored this VR environment. Moreover, I want to thank my friends Marc Santolini and Juan Giraldo for countless nights where they tried some new functionality, even though biology was not their area of expertise.

6.1 Introduction

In Chapter 5, I generated 4D datasets of developing embryos. Few software packages allow the import and observation of such large datasets. The tool I used throughout my thesis is BigDataViewer¹, a FIJI plugin developed for the express purpose of observing very large 4D microscopy datasets. However, despite its ease of use, BigDataViewer only allows the dynamic observation of slices through the dataset, and does not offer a reconstructed 3D view of the object. The software package Blender² allows users to observe 3D images, but does not support the fourth dimension of time. When it came to tracking nuclei, having to follow objects in 3D by observing a 2D slice was quite arduous and took a long time in Blender. Finally, all of these software packages project a three dimensional object onto a 2D image, approximating the depth information via perspective rules. Therefore, I aimed to build an alternative solution to observe 4D datasets.

I wanted this tool to solve multiple issues. First it needed to render the object in 3D and not as a 2D projection. Second, it needed to allow for the time axis such that 4D datasets could be loaded. Third, it needed to allow image parameter tuning, such as contrast and intensity. Fourth, it needed to be able to support the placements of annotations in 3D space to simplify tracking. I decided to use Virtual Reality as the underlying technology to address these issues. This chapter describes the advances that I made towards creating a tool to achieve these goals. The development of this tool relies on the previous work of many researchers to create hardware and software solutions tailored to the rendering of datasets. To place my work in this context, I first start by giving a short introduction to the history of computer rendering and

virtual reality. I then present the challenges of 3D images and the rendering methodology I chose to use. Finally I present the current context of VR development for microscopy analysis.

6.1.1 The many ways computers render images

With the advent of personal computers, one of the early artistic fields of programming was computer graphics (examples such as Thompson³ and reviewed in Tom⁴). Two main groups of people used this technology, the movie industry, and the hacker cracker scene (discussed in Silvast and Reunanen⁵). Starting in the 1970s, both groups started to create increasingly complex and artistic images (discussed in Silvast and Reunanen⁵). To this day, I believe that there remains a nostalgia of a time in the field of computer graphics that was governed by the rise of the demoscene⁵, an underground hacker artistic movement consisting of creating stunning computer graphics animations that would push beyond the foreseen limits of the available hardware of their time^{5,6}. Programmers would show their skills, coding software as small as a few kilobytes of code, creating stunning 3D artistic universes (discussed in Silvast and Reunanen⁵). Up until the end of the 1990s, each program had to use the programming language integrated into the graphical processing unit (GPU) (reviewed by Blythe⁷) and each device had a different access point interface (API) (reviewed by Blythe⁷). This changed with the invention in the 1980s⁸ of GPUs capable of executing a new, foreign set of instructions called a shader (the idea for shaders is credited to Cook⁸ and discussed in Hanrahan and Lawson⁹). With shaders, it was possible to execute an arbitrary set of logic rules, opening the doors for artists to experiment at will. First with the invention of RenderMan in 1988 by Pixar¹⁰, and then democratized by Nvidia in 2001 with the Geforce 3 GPU¹¹, shaders became the central pillar of modern computer graphics⁷. Today many shading languages exist and are used by different software packages and for different applications (discussed and reviewed in \citet{¹²}), but one of the most commonly used languages is the OpenGL Shading Language (GLSL)^{13,14}. While the demoscene still exists, the shading scene developed the concept of programmed computer graphics further. Today Visual DJs and digital artists use the power of shaders to generate stunning graphics¹⁵.

6.1.2 A short history of Virtual Reality and real-time rendering

Using the power of shaders, developers of game engines built the first tools allowing the creation and rendering of complex 3D graphics in Real-Time^{16,17}. The advent of real-time rendering opened the doors for today's vast array of games, simulations, data visualization tools, and much more. Multiple game engines were invented, each with different strengths and weaknesses, and each using different programming paradigms (reviewed by Paul et al.¹⁸). The most popular ones (in terms of numbers of game created using said engine^{19,20}) are Unreal Engine²¹ and Unity²², along with the rising open-source engine Godot²³. Those engines allow for the creation of customized tools such as the one presented here and custom rendering through the creation of shaders^{21,22}. This capacity to program unique behavior using the GPU compute potential is what allowed me to create the pilot Virtual Reality (VR) tool that I present in this chapter.

Here I define Virtual Reality as a technology whereby the user is immersed in a non-reality-bound universe, often 3D rendered. The user's perception is tricked into believing that what their eyes are seeing is real, even if they are actually a computer-generated projection. I would argue that the first step toward the invention of Virtual Reality started with the creation of the stereoscope in 1838 by Sir Charles Wheatstone²⁴. Using the concept of stereoscopy, the cinematographer Morton Heilig created Sensorama in the 1960s²⁵, a virtual reality booth offering viewers a 3D stereoscopic movie with vibration, scents, sound, and wind (Figure 6.1). He later invented the Telesphere Mask, what we would today refer to as a Virtual Reality Headset or a Head-Mounted Display (HMD)²⁶ (Figure 6.1). However, this invention lacked a critical aspect for the immersion of the user, namely the motion tracking of the head. In 1965, Ivan Sutherland invented another HMD that did track the user's head, linking it to computer rendering software that changed the position of the virtual camera to fully immerse the user in a virtual world²⁷. He called this display the Ultimate Display^{27,28} (Figure 6.1), and from this technology emerged what we know of today as Virtual Reality (reviewed by Mandal²⁹) (Figure 6.1). Until the creation of the Oculus headset in 2012 (patented in 2014³⁰), HMD remained bulky, expensive, and ill-equipped tools (reviewed by Mandal²⁹). At the time of its creation, the first Oculus HMD could only track the user's head in three degrees of freedom (3

DoF) (Figure 6.1), therefore the user's position in space was not registered by the computer, and the position of the virtual camera was not updated³⁰. Later iterations by Oculus, and then by HTC³¹ and Valve³² allowed for the creation of the fully immersive HMD now commercially available (Figure 6.1). Those headsets all possess 6 DoF, tracking not only the user's head direction but also its position (Figure 6.1). The addition of hand controllers permitted the projection of the user's hands in the simulation, increasing the immersion, and allowing for direct interactions with virtual objects (Figure 6.1).

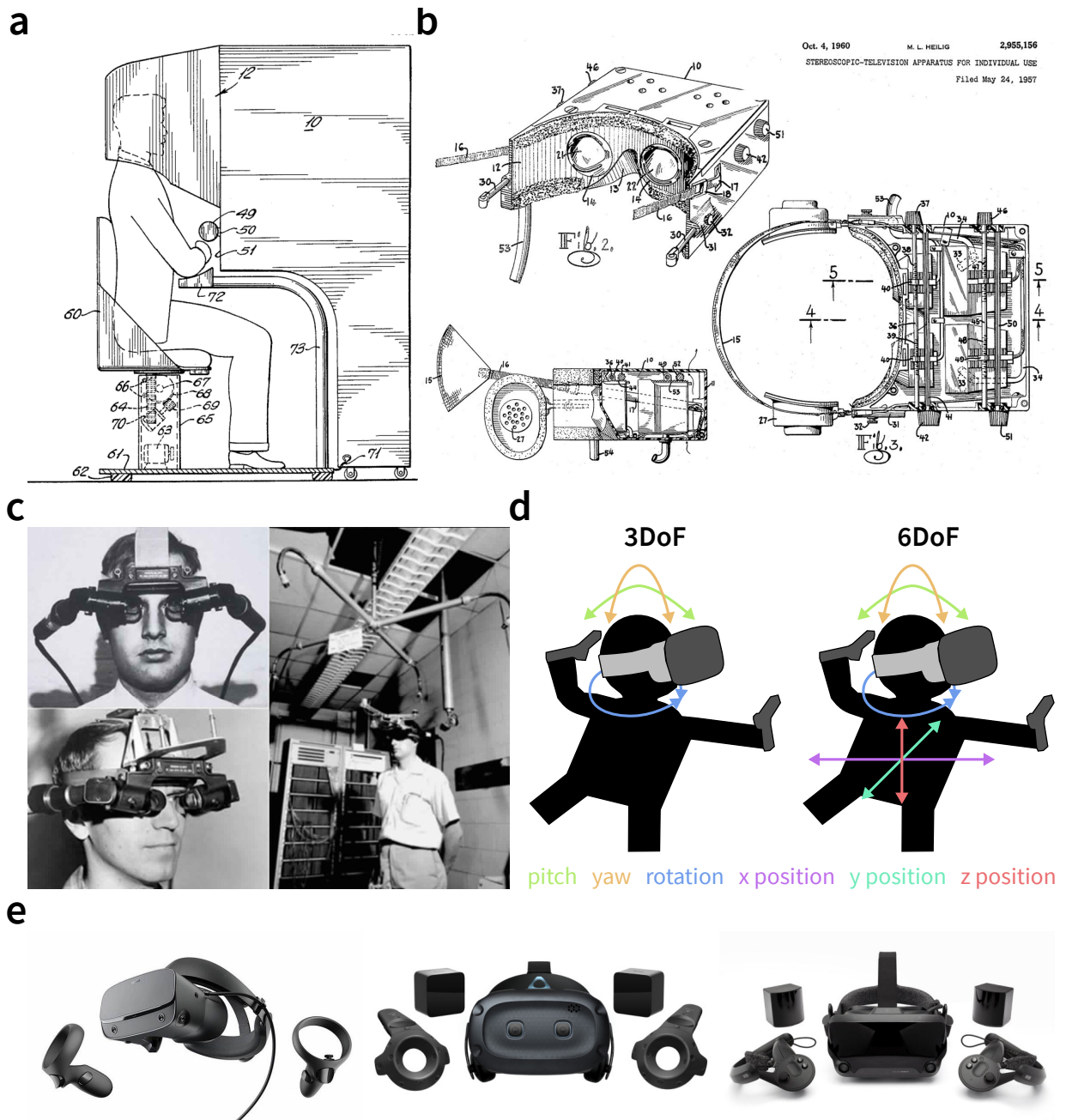


Figure 6.1: History of VR headset technological development. **a)** First virtual reality cinema, the Sensorama created in 1956 by Morton Heilig (image from Heilig²⁵). **b)** First patented design for a virtual reality HMD designed by Morton Heilig, the Telesphere (image from Heilig²⁶). **c)** Photographs of the Ultimate Display system created by Ivan Sutherland (images from the original paper by Sutherland²⁸) **d)** Schematic representation of three and six degrees of freedom (DoF) applied to virtual reality motion tracking. **e)** Example of modern virtual reality headsets, from left to right Oculus Rift, HTC Vive, Valve Index.

6.1.3 Rendering 3D images and the implications for light sheet data

I generated the datasets I used to develop the software I present here with light sheet microscopy, which was introduced at length in Chapter 5 of this thesis. After processing those microscopy datasets, the final result is a volumetric image, defined as a three-dimensional image composed of voxels whose positions are defined by three orthogonal euclidian axes, here called x , y , and z (Figure 6.2). Contrary to a two-dimensional image (Figure 6.2), a volumetric image cannot be directly rendered onto a 2D computer screen without being projected to two dimensions. The stereoscopic nature of Virtual Reality makes it possible to render such an image in three dimensions, each eye is only observing a two-dimensional projection of the scene, shifted slightly such that the brain reconstructs it into three dimensions (Figure 6.2).

A rendering technique is a set of instructions that will take the elements in a virtual environment, and generate an image that can be visualized on a screen. One technique to render volumetric images is called Ray Marching³³. For each pixel to be rendered, one virtual ray will be projected from the virtual camera, with position O (for origin) and in the direction \vec{D} (Figure 6.3). If O is a point in space and \vec{D} a normalized vector, then we can define any point along that line by $P = O + t * \vec{D}$, where t is a scalar corresponding to the distance along the ray to the origin (Figure 6.3). To sample our 3D image, for each pixel, the shader computes the sum of intensity values along that ray at multiple positions P , where at each step we increase t with a small value dt (Figure 6.3). When we reach the end of the 3D image we stop and we return the total value for that pixel. The resulting image will be a sum of intensity projections of our 3D image onto a 2D image (Figure 6.3). One can then change this shader, for example, to return the maximum value, or the minimum value, or any other mathematical function. Furthermore, the values returned can then be blended in with a Look-Up Table (LUT), the intensity modified, or the values scaled non-linearly through a power law. Moreover, multiple performance improvements can be coded into this shader, such as an automatic loop break when the sum of intensity reaches its maximum allowed value of 1.0, which avoids extraneous dt steps. Another performance improvement is the breakdown of the volume into smaller cubes, where each cube that does not contain information is removed from the rendering pipeline. This latter technique can be further improved for performance by dynamically rendering (culling) the cubes facing

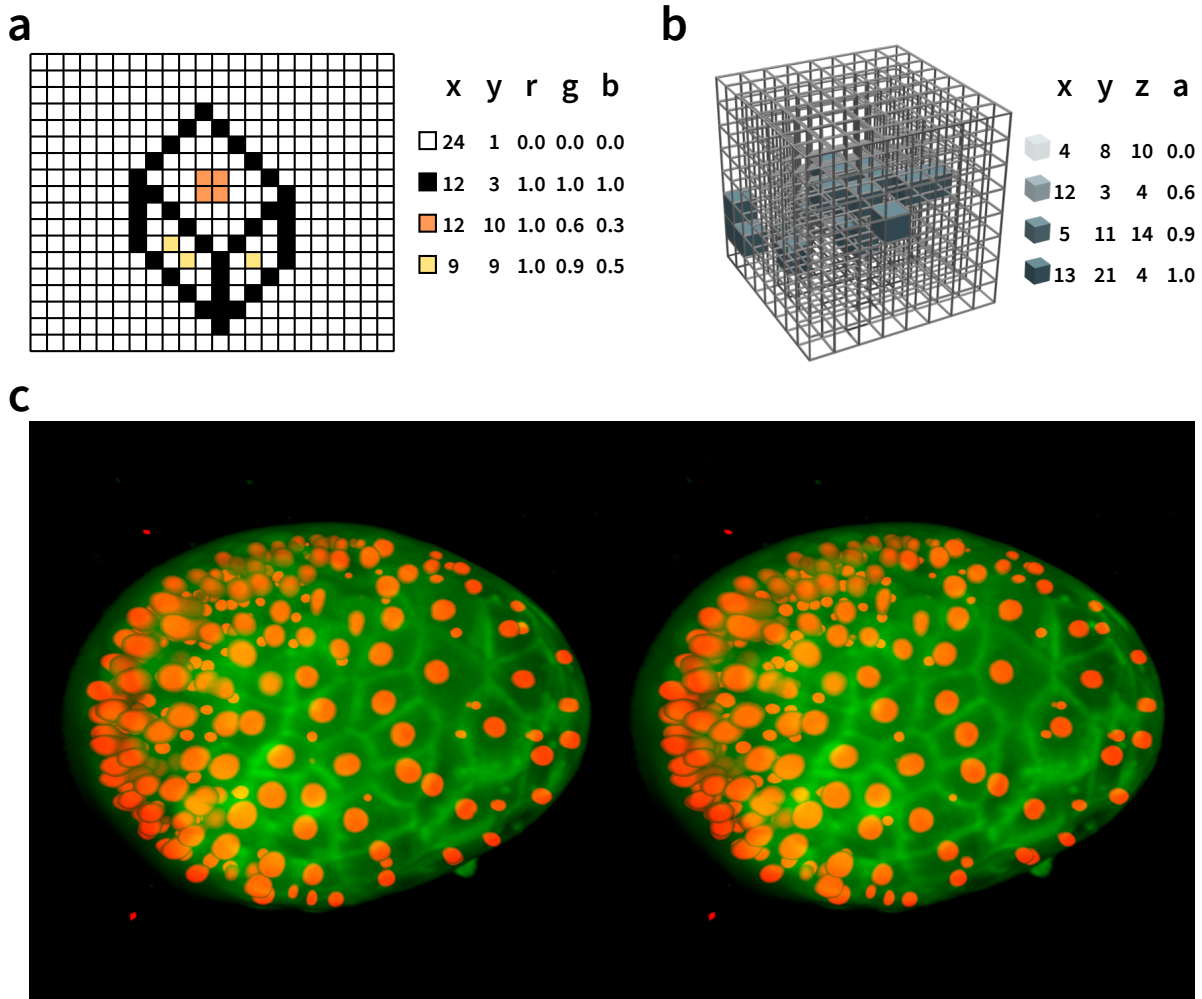


Figure 6.2: Construction of 2D, 3D images and stereoscopy. **a)** schematic representation of a 2D image. Each pixel is represented by a set of five values, its position in the grid as x and y , and color as red (r) green (g) and blue (b) values. **b)** Schematic representation of a 3D image. Each voxel is represented as a cube with a position in space given by x y and z as well as a transparency value α (a). **c)** Example of a stereoscopic rendering of a *P. hawaiiensis* embryo at stage S6 imaged with a light sheet microscope. In red nuclei are shown and in green the membranes are shown. The 3D image is rendered from two different camera points distanced by the average interpupillary distance to mimic three-dimensional vision. To see this object in 3D, place a sheet of paper between both images and look at them with the left and right eye only seeing the left or right image (a good tutorial on how to do this can be found here: <http://www.neilcreek.com/2008/02/28/how-to-see-3d-photos/>).

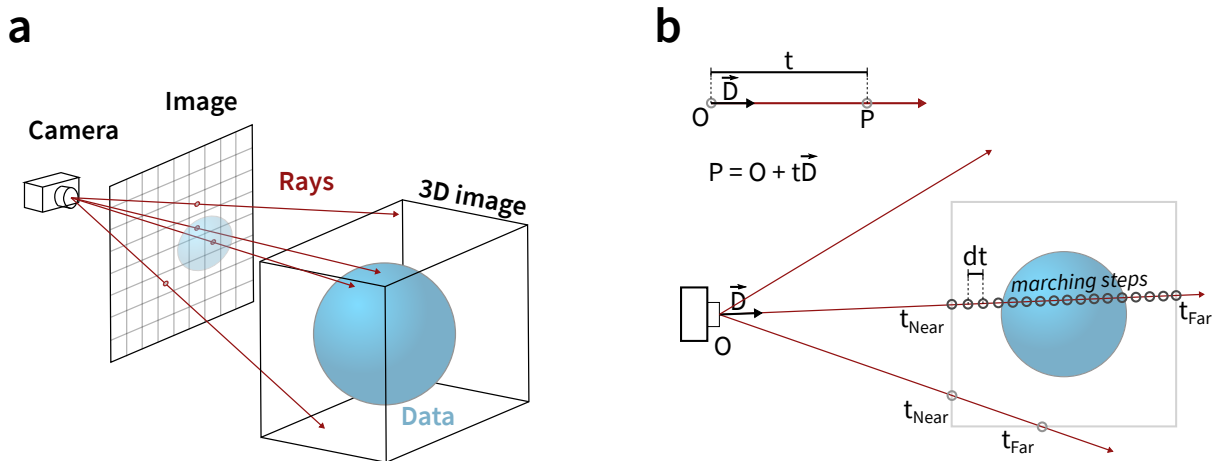


Figure 6.3: Schematic representation of Ray Marching. **a)** 3D schematic showing the projection of rays from the camera onto a 3D image. Each ray (in red) crosses the image plane at the center of each pixel then continues towards the 3D image. **b)** Schematic 2D projection of Ray Marching. Each ray is projected from the camera and may intersect the 3D image. In case it intersects, it will then sample the image along the ray every dt . The pixel value for that ray will be, for example, the sum of sampled values along the ray.

the camera. Finally, new rendering techniques, albeit more complex to implement, have been developed to allow the real time rendering of very large volumetric dataset^{34,35}. In this chapter, I will describe the coding of a primitive Ray Marching shader for Unity, which uses simple performance enhancement.

Finally, while this chapter focuses on the rendering of microscopy images, the software that I have begun to create is in principle not limited to the observation of 3D microphotographs, but rather should be able to load any volumetric image. Magnetic Resonance Imaging (reviewed in Williams and Drew³⁶) and Computed Tomography scans (reviewed in Williams and Drew³⁶) are two imaging techniques that generate a 3D image of the (often human) specimen. The medical field has been using specialized tools and software allowing for the rendering and analysis of such datasets for a long time (reviewed in El Beheiry et al.³⁷, Feng et al.³⁸). However, these tools are specialized for the medical community and do not often offer the adjustment or analysis needed for other types of biological samples. For example, MedicalHolodeck³⁹ offers a VR visualizer for MRI and scan data, DICOM VR⁴⁰ lets you import DICOM datasets (a common file format for medical 3D images) and annotate areas by painting over structures. However, at the time of writing, the tool was still under development and not yet publicly available⁴⁰.

Earlier work with 3D screens and CAVE systems had developed VR tools^{41,42,43}. When I started

this project, to my knowledge there existed three software that allowed a user to import 3D microscopy datasets in HMD based VR: Arivis Vision VR⁴⁴ and syGlass⁴⁵ which are commercial products, and an open source and peer reviewed tool tailored to 3D colocalization analysis in confocal images⁴⁶. Another study⁴⁷ reported the development of a tool to observe light sheet images but did not publish their code making it impossible to use or iterate on their research. Moreover, the use of VR was also developed for scientific communication such as Journey to the center of the cell⁴⁸ that used electron microscopy data of a cell to recreate a virtual cell. In between the time I started this project and the time of writing this thesis newer tools emerged such as ConfocalVR⁴⁹ which was developed to observe confocal 3D stacks in VR. Other microscopy techniques such as single molecule microscopy have also seen recent tools developed for such tasks⁵⁰.

Other aspects of biological systems are also starting to be integrated with VR systems such as the observation of DNA, RNA and protein structure in BioVR⁵¹, or the co-creation of molecular simulations in VR⁵². Connectomics approaches are also seeing new VR tools to observe the results of electron microscopy stacks and allow users to perform annotation or refine the segmentation⁵³. Finally, a great review detailing the use of VR for data visualization in biology has recently been published³⁷. Here, I present my work towards the creation of a VR software to enhance the perception of three-dimensional biological structures and the tracking of objects in 3D. The goal of this tool is to allow users to observe a 3D dataset and immerse themselves into, for example, a developing embryo. I aimed to allow users to use this tool to manipulate 3D data of biological tissues as a three-dimensional object, performing cuts, deformations, or annotations on the data directly in space, therefore in three dimensions. I (and others³⁷) believe that this technology revolutionizes our intuitive understanding of biological systems as it is now possible to observe the entirety of the data in an immersive way instead of a two-dimensional projection.

6.2 Methods

6.2.1 Development of the software

The software was developed in C# and Unity shader language for the Unity game engine⁵⁴. The SteamVR and VRTK libraries were used for locomotion and interactions. The TextMeshPro library was used for the UI panels. The current development requires a steamVR or openXR compatible VR HMD. The code is available at https://gitlab.com/xqua/microscopy_vr.

6.2.2 Ray Marching shader

The Ray Marching shader was coded in Unity shader language and can be found in the main repository: https://gitlab.com/xqua/microscopy_vr. The shader is customizable via the following exposed parameters:

- **_Data:** Data Texture -> Texture3D
- **_DataChannel:** Data Channel for monochromatic dataset, normally use A8 so 4th channel => Vector4(0,0,0,1)
- **_Axis:** Axes order -> Vector3(1, 2, 3)
- **_TexFilling:** Data filling factors -> Vector3(1.0, 1.0, 1.0)
- **_Color:** Color for shading monochromatic dataset -> Vector3(1.0, 1.0, 1.0)
- **_SliceAxis1Min:** Slice along axis X: min -> Range(0, 1)
- **_SliceAxis1Max:** Slice along axis X: max -> Range(0, 1)
- **_SliceAxis2Min:** Slice along axis Y: min -> Range(0, 1)
- **_SliceAxis2Max:** Slice along axis Y: max -> Range(0, 1)
- **_SliceAxis3Min:** Slice along axis Z: min -> Range(0, 1)
- **_SliceAxis3Max:** Slice along axis Z: max -> Range(0, 1)
- **_C1:** Channel 1 ON/OFF -> Range(0,1)

- **_C2:** Channel 2 ON/OFF -> Range(0,1)
- **_C3:** Channel 3 ON/OFF -> Range(0,1)
- **_RGB:** Texture type: A8 = 0, RGBA32 = 1 -> Float
- **_DataMinR:** R: Data threshold: min -> Range(0, 1)
- **_DataMaxR:** R: Data threshold: max -> Range(0, 1)
- **_DataMinG:** G: Data threshold: min -> Range(0, 1)
- **_DataMaxG:** G: Data threshold: max -> Range(0, 1)
- **_DataMinB:** B: Data threshold: min -> Range(0, 1)
- **_DataMaxB:** B: Data threshold: max -> Range(0, 1)
- **_StretchPowerR:** R: Data stretch power -> Range(0.1, 3)
- **_StretchPowerG:** G: Data stretch power -> Range(0.1, 3)
- **_StretchPowerB:** B: Data stretch power -> Range(0.1, 3)
- **_NormPerStepR:** R: Intensity normalization per step -> Float
- **_NormPerStepG:** G: Intensity normalization per step -> Float
- **_NormPerStepB:** B: Intensity normalization per step -> Float
- **_NormPerRay:** Intensity normalization per ray -> Float
- **_Steps:** Max number of steps -> Range(1,1024)

6.2.3 Conversion of BigDataViewer HDF5 files to RAW

The conversion of BigDataViewer HDF5 files onto RAW and JSON files was done using a custom python script. This script can be found in the [LightSheetUtils repository](#). The parameters are Options:

-h, --help show this help message and exit

-i FILENAME, --input=FILENAME [REQUIRED] hdf5 file path

-o FILENAMEOUT, --output=FILENAMEOUT [REQUIRED] bin file path

-l LEVEL, --level=LEVEL [REQUIRED] resolution level to extract (0: full size, 1: 2x downsample, 2: 4x downsample, 3: 8x downsample)

-o To, --timepoint-start=To first time point to extract

-n TN, --timepoint-end=TN last time point to extract

-C, --multichannel Is the dataset a multichannel dataset?

-N, --normalize Global intensity normalization?

-c CROP, --crop=CROP Number of pixels to crop from the cube

-m MIN, --min=MIN Manually set the min. channel1,channel2

-M MAX, --max=MAX Manually set the max. Channel1,channel2

6.3 Results

6.3.1 Observing 3D images and coding a Ray Marching shader

I used Unity game engine to develop a Ray Marching shader to observe the *P. hawaiiensis* light sheet data gathered in Chapter 5 in 3D. Tools to make the development of virtual reality applications had already been developed by others, such as the open source library Virtual Reality ToolKit (VRTK)⁵⁵. Because 3D images are by nature a cuboid (three-dimensional rectangle), I needed to render them into a cuboidal mesh. Therefore the shader must project the 3D image onto the faces of that cube. To this end, the shader I coded needed to first detect the bounds of the cuboid, then the position of the faces so that the algorithm could start the ray marching steps from that point forward (Figure 6.3). This is achieved using the function `IntersectBox` shown in Listing 1:

Listing 1: Hitting the bounding box. This function takes a ray origin, a ray direction, the bounding box and returns a boolean that is true if the ray hits the bounding box. It also returns the closests and farthest points.

```
bool IntersectBox(float3 ray_o, float3 ray_d, float3 boxMin,
                  float3 boxMax, out float tNear, out float tFar)
{
    // compute intersection of ray with all six bbox planes
    float3 invR = 1.0 / ray_d;
    float3 tBot = invR * (boxMin.xyz - ray_o);
    float3 tTop = invR * (boxMax.xyz - ray_o);
    // re-order intersections to find smallest and largest on each axis
    float3 tMin = min (tTop, tBot);
    float3 tMax = max (tTop, tBot);
    // find the largest tMin and the smallest tMax
    float2 t0 = max (tMin.xx, tMin.yz);
    float largest_tMin = max (t0.x, t0.y);
    t0 = min (tMax.xx, tMax.yz);
    float smallest_tMax = min (t0.x, t0.y);
    // check for hit
    bool hit = (largest_tMin <= smallest_tMax);
    tNear = largest_tMin;
    tFar = smallest_tMax;
    return hit;
}
```

Listing 1 returns information on whether a ray from the camera hit an object, along with the two intersection distances t_{Near} and t_{Far} for that ray (as we have seen in the introduction, those t values correspond to a unique point P on the ray) (Figure 6.3). The Ray Marching loop is then

started from the originating point t_{Near} , then stops when we either hit our maximum saturation value of 1.0 for that pixel or when we hit the back of the cube t_{Far} . To perform the loop, a set dt is added to t_{Near} and the shader samples the 3D image at each new point P . Listing 2 shows the monochrome sampling function `get_data`. The pixel value is then colorized by being multiplied by a set color.

Listing 2: Sampling data from the 3D image This function takes a position in the 3D image, samples the intensity value for that given voxel, and returns a color vector.

```
bool IntersectBox(float3 ray_o, float3 ray_d, float3 boxMin,
                 float3 boxMax, out float tNear, out float tFar)
{
    ...
}

// gets data value at a given position
float4 get_data(float3 pos) {
    // sample texture (pos is normalized in [0,1])
    // This sets the axis to read the texture, it allows to
    // change the order in which the texture is Loaded
    // For example it can reverse an axis or switch them
    float3 posTex = float3(pos[_Axis[0]-1],
                          pos[_Axis[1]-1],
                          pos[_Axis[2]-1]);
    // This allows to display dataset that are not isomorphic
    // at a pseudo isomorphic dataset
    // _TexFilling is a vec3 of scaling factors, for example
    // if the resolution is x: 0.5um y: 0.5um z: 2um then
    // a _TexFilling of vec3(1, 1, 0.25) should be used
    posTex = (posTex-0.5) * _TexFilling + 0.5;
    // This loads the value at the position posTex with the
    // Level of Detail 0, aka with no MipMapping (so that
    // we get an accurate reading and not an interpolation)
    float4 data4 = tex3Dlod(_Data, float4(posTex,0));

    // colourize
    float4 col = float4(data4.a, data4.a, data4.a, data4.a);
    col *= _Color;
    return col;
}

float4 frag(frag_input i) : COLOR
{
    ...
}
```

Finally, we can write the fragment shader function (also called pixel shader). Listing 3 is the main loop that will be executed for each pixel in the rendering pipeline. We obtain the starting and stopping positions with the `IntersectBox` function described above, and loop for a set

number of values through the 3D image. The number of steps `_Steps` sets the distance between each sampling point along that ray dt . Finally, we blend the value obtained starting from the closest point to the camera and giving less weight to values from further inside the volume. This guarantees that the contribution of data closest to the camera will get precedence over voxels that are behind them in the same ray trajectory. Listing 3 shows the complete fragment function. This code was implemented as a Unity shader and tested on a 3D image generated in Chapter 5. The result is shown in Figure 6.4.

Listing 3: The fragment shader main function. This is the main function of the shader, it computes the ray origin and direction, then checks if that ray hits the bounding box of our 3D image by calling `IntersectBox`. If the ray hits the box, it performs the Ray Marching steps by adding the intensity values sampled from the `get_data` function. Once it reaches the end of the box, or saturates at 1.0 the loop is broken and the color for this pixel is returned.

```
float4 frag(frag_input i) : COLOR
{
    i.ray_d = normalize(i.ray_d);
    // calculate eye ray intersection with cube bounding box
    // The cube is defined as a unit cube centered around the origin
    float3 boxMin = { -0.5, -0.5, -0.5 };
    float3 boxMax = { 0.5, 0.5, 0.5 };
    // check if the ray intersects the bounding box
    // This will hold the t values for the near and far faces
    float tNear, tFar;
    // run the intersection function to check if the ray hits
    bool hit = IntersectBox(i.ray_o, i.ray_d, boxMin, boxMax,
        tNear, tFar);
    if (!hit) discard; // If it does not hit, stop the processing
    // If the ray hits but that the nearest t is negative then it
    // is behind us, therefore set it to 0 to start the ray
    // marching from that point in space
    if (tNear < 0.0) tNear = 0.0;
    // calculate intersection points
    float3 pNear = i.ray_o + i.ray_d*tNear;
    float3 pFar = i.ray_o + i.ray_d*tFar;
    // convert to texture space, we add 0.5 because our unit cube
    // is centered around the origin. therefore the positions
    // from -0.5 to 0.5 must be converted into 0 to 1.
    pNear = pNear + 0.5;
    pFar = pFar + 0.5;

    // march along ray inside the cube, accumulating color
    // First set the starting point for sampling as the nearest
    // intersection
    float3 ray_pos = pNear;
    // Then set the vector direction as the direction between
    // the nearest and the far intersects
    float3 ray_dir = pFar - pNear;
    // We then create dt as a small vector that we can add to
```

```

// our sampling point
// We set the _Steps as a value which corresponds to the
// number of steps taken to sample the volume
float3 ray_step = normalize(ray_dir) * sqrt(3) / _Steps;
// We initialize the color of the pixel
float4 ray_col = 0;

// Start the main marching loop between 0 and 1 with _Steps
for(int k = 0; k < _Steps; k++)
{
    // We sample the 3D image at that point
    float4 voxel_col = get_data(ray_pos);

    // Then we blend in the value sampled with the previous values
    // This blending function gives more weight to voxel that
    // are near the camera
    // and decreases the value as we move further inside the cube
    ray_col = ray_col + (1 - ray_col) * voxel_col;

    // We move the sampling point by dt
    ray_pos += ray_step;

    // Then we check that we are still within the cube otherwise
    // we will sample outside the 3D image
    if (ray_pos.x < 0 || ray_pos.y < 0 || ray_pos.z < 0) break;

    if (ray_pos.x > 1 || ray_pos.y > 1 || ray_pos.z > 1) break;

    // If the pixel value for this ray is already saturated,
    // break the loop.
    if (ray_col.a > 1.0) break;
}
// Clamp the value between 0 and 1 to avoid erroneous colors
ray_col = clamp(ray_col,0,1);
// Finally return the color for that pixel.
return ray_col;
}

```

To allow for the modifications of contrasts and intensities in a similar fashion as by software such as FIJI⁵⁶, I modified the shader to perform voxel normalization and intensity clamping. First I added a minimum intensity threshold (`_DataMinR`, whereby any value in the image lower than this would return 0. Similarly, I added a maximum value threshold (`_DataMaxR`) which clamped any voxel above a certain value to 1. Second, I added a normalization to the returned color that multiplies that voxel value by a step normalization factor (`_NormPerStep`). Finally, I added a data stretching normalization by taking the power of the voxel value (`_StretchPower`). This allows the user to fine-tune each unique dataset, and maximizes the rendering quality, along with letting the user choose which part of the sample needs to be

displayed at any given time. The results of a fine-tuned rendering can be seen in Figure 6.4.

Listing 4: Intensity normalization. The `get_data` function is modified to clip values with a lower and upper bound. The ray marching loop normalizes the sampled intensity value by taking that value to the power of `_StretchPower`.

```
bool IntersectBox(float3 ray_o, float3 ray_d, float3 boxMin,
                 float3 boxMax, out float tNear, out float tFar)
{
    ...
}

// gets data value at a given position
float4 get_data(float3 pos) {
    ...
    // We pass the value for the voxel through a step function
    // which returns 0 if the value is below the first argument
    // And returns 1 if the value is above the second argument
    data4.a *= step(_DataMinR, data4.a);
    data4.a *= step(data4.a, _DataMaxR);
    ...
}

float4 frag(frag_input i) : COLOR
{
    ...
    // Start the main marching loop between 0 and 1 with _Steps
    for(int k = 0; k < _Steps; k++)
    {
        // We sample the 3D image at that point
        float4 voxel_col = get_data(ray_pos);
        //
        voxel_col.a = _NormPerStep
                    * length(ray_step)
                    * pow(voxel_col.a, _StretchPower);
        ...
    }
    ...
}
```

I added the capability to slice the dataset along any axis in order to expose the inside of the image. This was implemented by modifying the sampling function and checking whether the position sampled is before or after the cutting plane. New parameters (`_SliceAxisMin` and `_SliceAxisMax`) to set the min and max planes were added and the resulting code can be seen in the Listing 5. The results can be seen in Figure 6.4.

Listing 5: Slicing the 3D image. To allow for slicing of the 3D image along the three axes, the `get_data` function is modified. For each axis x , y and z the returned value is clamped to 0 if the position is beyond the slicing parameter for this axis.

```
bool IntersectBox(float3 ray_o, float3 ray_d, float3 boxMin,
                 float3 boxMax, out float tNear, out float tFar)
{
    ...
}

// gets data value at a given position
float4 get_data(float3 pos) {
    ...
    // We check that the sampled position is within
    // the bounds set by the Slice Axis parameters
    data4 *= step(_SliceAxis1Min, posTex.x);
    data4 *= step(_SliceAxis2Min, posTex.y);
    data4 *= step(_SliceAxis3Min, posTex.z);
    data4 *= step(posTex.x, _SliceAxis1Max);
    data4 *= step(posTex.y, _SliceAxis2Max);
    data4 *= step(posTex.z, _SliceAxis3Max);
    ...
}

float4 frag(frag_input i) : COLOR
{
    ...
}
```

Finally, fluorescent microscopy today allows the recording of multiple fluorophores in the same sample (as in Chapter 5). In my case, *P. hawaiiensis* embryos expressed H2A-mCherry and Lyn-GFP, markers of nuclei and membranes respectively. I wanted to display the colors for each of the two channels onto the rendered 3D image. To this end, I changed the image format in the shader from the monochromatic A8 (alpha eight bit) to the trichromatic RGB8 (red green blue eight bit). Then at each step, I sample each of the three colors and keep them separate as RGB values. Finally, each channel was given its own normalization, clipping, and stretching factors for fine-tuning each color. The resulting code can be seen in Listing 6. The complete final shader for the Unity engine can be found in the Methods section: *Ray Marching shader*.

Listing 6: Displaying multiple channels. To allow the visualization of up to three microscopy channels the texture type was modified from Alpha8 to RGB8. Each channel is sampled accordingly in the `get_data` function and the normalization steps are modified to be applied to each channel. Each channel uses a different set of normalization parameters that allows fine tuning of the image normalization.

```
// gets data value at a given position
float4 get_data(float3 pos) {
    ...
    float4 data4 = tex3Dlod(_Data, float4(posTex,0));
    ... // Plane clipping is done here
    data4.r *= step(_DataMinR, data4.r);
    data4.r *= step(data4.r, _DataMaxR);
    data4.g *= step(_DataMinG, data4.g);
    data4.g *= step(data4.g, _DataMaxG);
    data4.b *= step(_DataMinB, data4.b);
    data4.b *= step(data4.b, _DataMaxB);

    float4 col = float4(data4.r,
                       data4.g,
                       data4.b,
                       data4.r+data4.g+data4.b);
    // We can activate or deactivate a channel through
    // a change in parameter.
    if (_C1 == 0) {
        col.r = 0;
    }
    if (_C2 == 0) {
        col.g = 0;
    }
    if (_C3 == 0) {
        col.b = 0;
    }
    return saturate(col);
}

float4 frag(frag_input i) : COLOR
{
    ...
    float4 voxel_col = get_data(ray_pos);
    // Each channel gets its own normalization factor.
    voxel_col.r = _NormPerStepR
                 * length(ray_step)
                 * pow(voxel_col.r, _StretchPowerR);
    voxel_col.g = _NormPerStepG
                 * length(ray_step)
                 * pow(voxel_col.g, _StretchPowerG);
    voxel_col.b = _NormPerStepB
                 * length(ray_step)
                 * pow(voxel_col.b, _StretchPowerB);
    ...
}
```

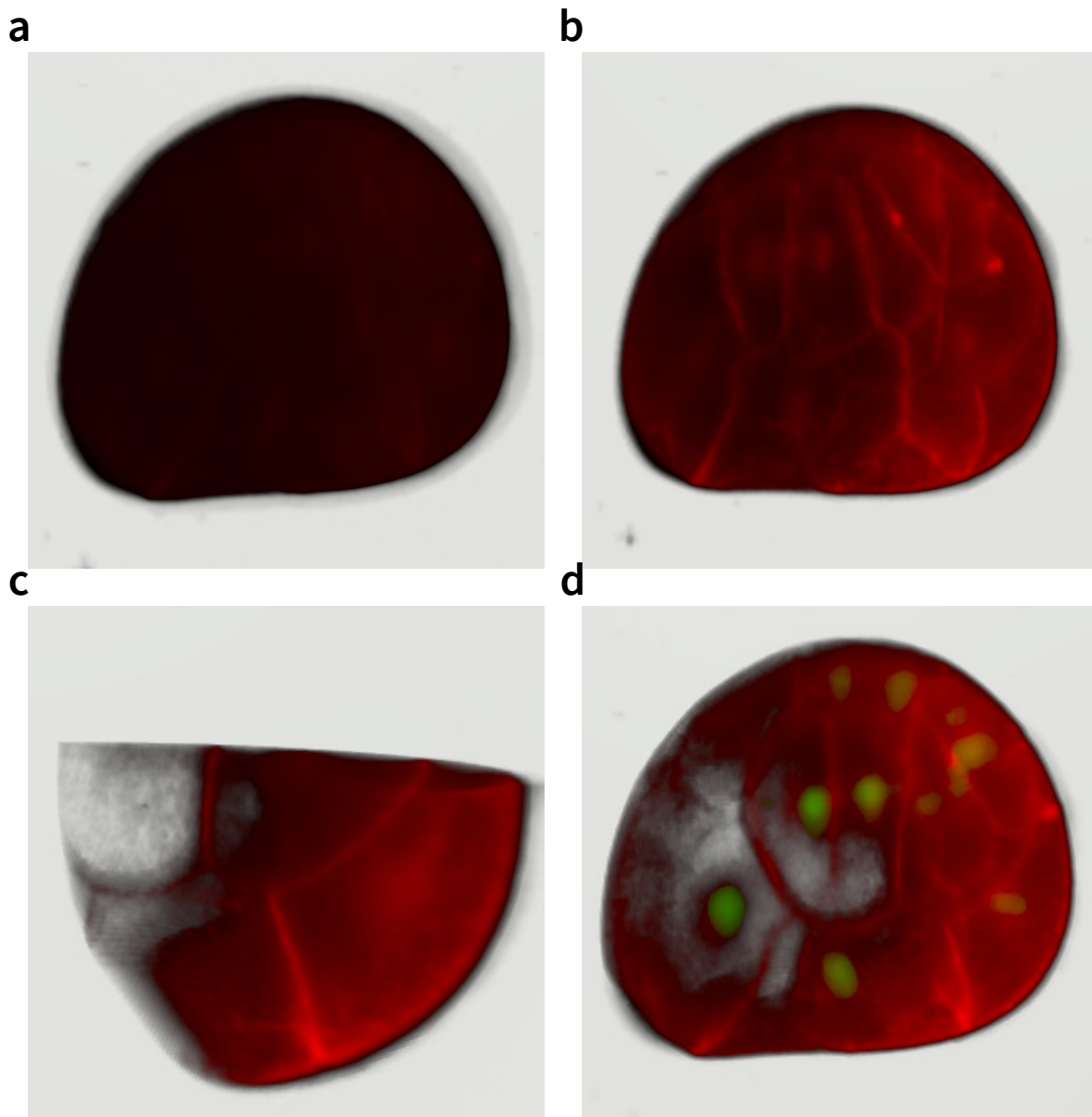



Figure 6.4: Example images of *P. hawaiiensis* embryos at the 16 cell stage (stage S5) resulting from the different iterations of the volume rendering shader. a) Rendering with no normalization or stretching applied to the 3D image. b) Rendering with normalization and stretching applied to the 3D image. c) Rendering of the volume sliced on the Y and Z axis to reveal the interior of the sample. d) Rendering of the volume with multiple microscopy channels represented as red and green values.

6.3.2 Dataset handling strategies

Having developed a working shader, the next requirement was for the user to be able to load and manage datasets. Contrary to 2D images, where file formats have been stable for years⁵⁷ (e.g. JPG, PNG, TIFF), 3D image formats do not have unified standards despite efforts to reach a consensus⁵⁷. For example, Ilastik⁵⁸ and BigDataViewer¹, which were used in Chapter 5, both store their data in an HDF5 database. However, their data storage strategies are completely different. Ilastik stores the volumetric data as multi-layer 2D images⁵⁸, whereas BigDataViewer stores chunks of the volume in smaller cubes (called cube chunks)¹, as well as downsampled versions of the data for fast access in real-time rendering (Figure 6.5). Another format for 3D images is the RAW format. It is a linear string of bytes, and requires that metadata about the 3D image be stored and accessed from another location. Due to the efficiency of the BigDataViewer data handling, I aimed to load the 3D images from HDF5 files using the BigDataViewer structure. Under this structure, all the necessary metadata is written to an XML file and the setting for each dataset can then be written to that file and recovered in later sessions. However, the Unity C# environment did not come with a functional HDF5 library. Moreover, after many attempts, I failed at compiling a native version of the C++ HDF5 library for Unity. Therefore, I set aside the goal of reading from HDF5 files of volumetric data and instead used the much simpler RAW format.

To this end, I created a JSON file holding the metadata for each dataset, namely the size of the cube in x y and z, the number of channels, and the number of time points. Finally, I created a small python script that would convert a BigDataViewer HDF5 file into a set of RAW files with the attached JSON metadata at the desired resolution (this script is available at: https://github.com/extavourlab/Blondel_Leo_Thesis).

I then implemented a database manager in Unity. For each dataset, it reads the JSON file, and populates the database accordingly. For each dataset, a maximum number of allowed bytes is set. This was set to 4 Gbytes due to the memory limitation of the GPU I used to create this software, but can be changed for different hardware. Finally, I preloaded the dataset into the GPU memory via the instantiation of a *Texture3D* array. This pre-instantiation is necessary to be able to smoothly play a 3D image movie.

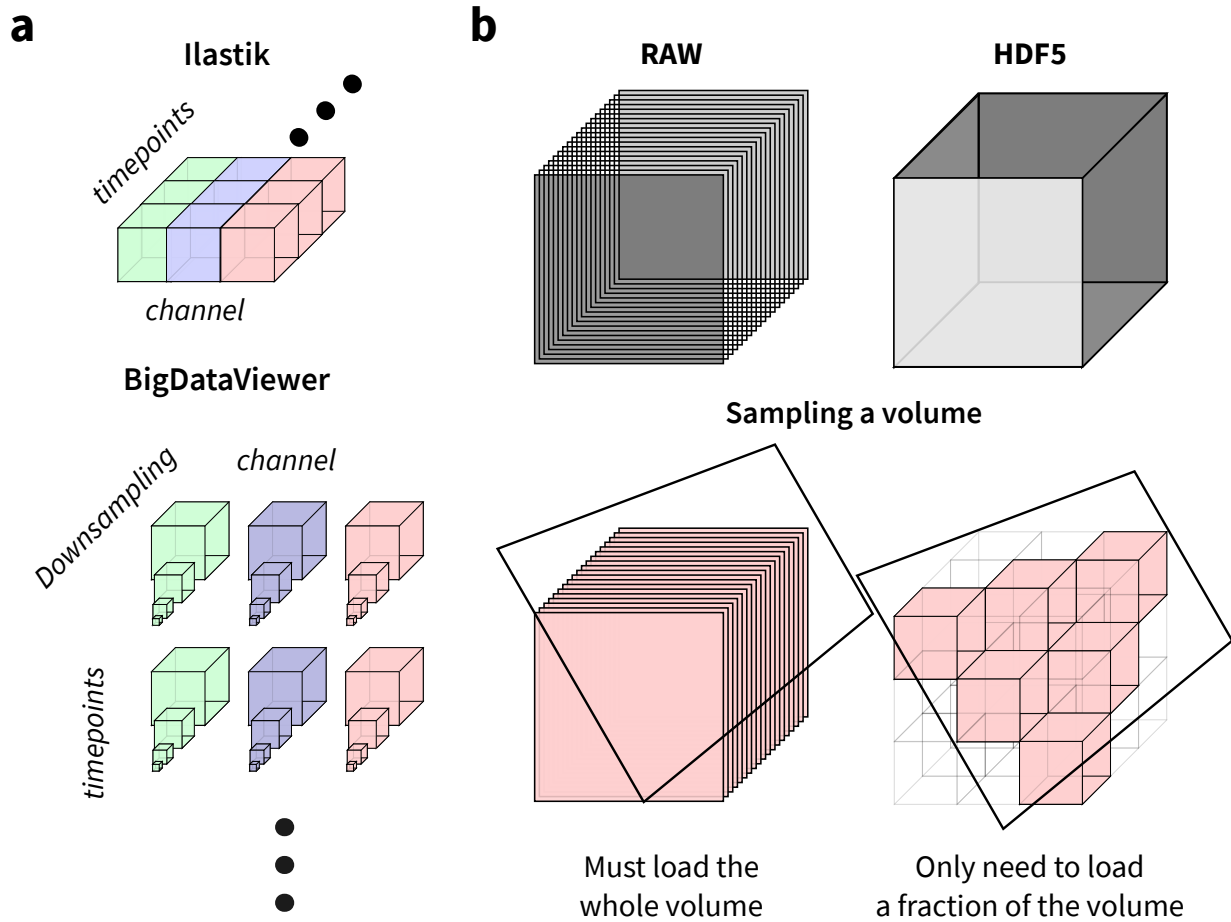


Figure 6.5: Schematic representation of different 3D image file format strategies. **a)** Schematic representation of the storing strategy in an HDF5 database for Ilastik and BigDataViewer. Ilastik stores the dataset in two dimensions within the HDF5 database, one for the time point and one for the channel⁵⁸. BigDataViewer takes advantage of the HDF5 format by adding a third dimension which are downsampled versions of the main dataset for faster access¹. **b)** Example of RAW versus HDF5 for storing a 3D image. In a RAW formatted file, the entire data must be loaded into memory to access a point within the 3D space. In HDF5, thanks to the chunking of the data, it is possible to retrieve specific parts of the 3D image. Each smaller part is orders of magnitude faster to load than the whole dataset¹. This is a key feature for performance optimization.

6.3.3 Designing the user experience for manipulating 3D images in VR

The next step was to construct a user interface to interact with the data. Doing this poses new challenges, as the world that the user is projected into is itself the design object that must be created. Classical interfaces such as UI menus, actions, and mouse interactions do not apply in VR. Therefore, I decided on a few parameters that I considered essential to the experience. First, the user must be able to move around the space to observe the dataset from any angle. Second, the user must be able to load multiple datasets in the same room to compare them. Third, the user must be able to tune each dataset parameter and perform different analyses as intuitively as possible, without occluding the view of the dataset. Those ideas can be seen in the blueprint design I created before starting the implementation (Figure 6.6).

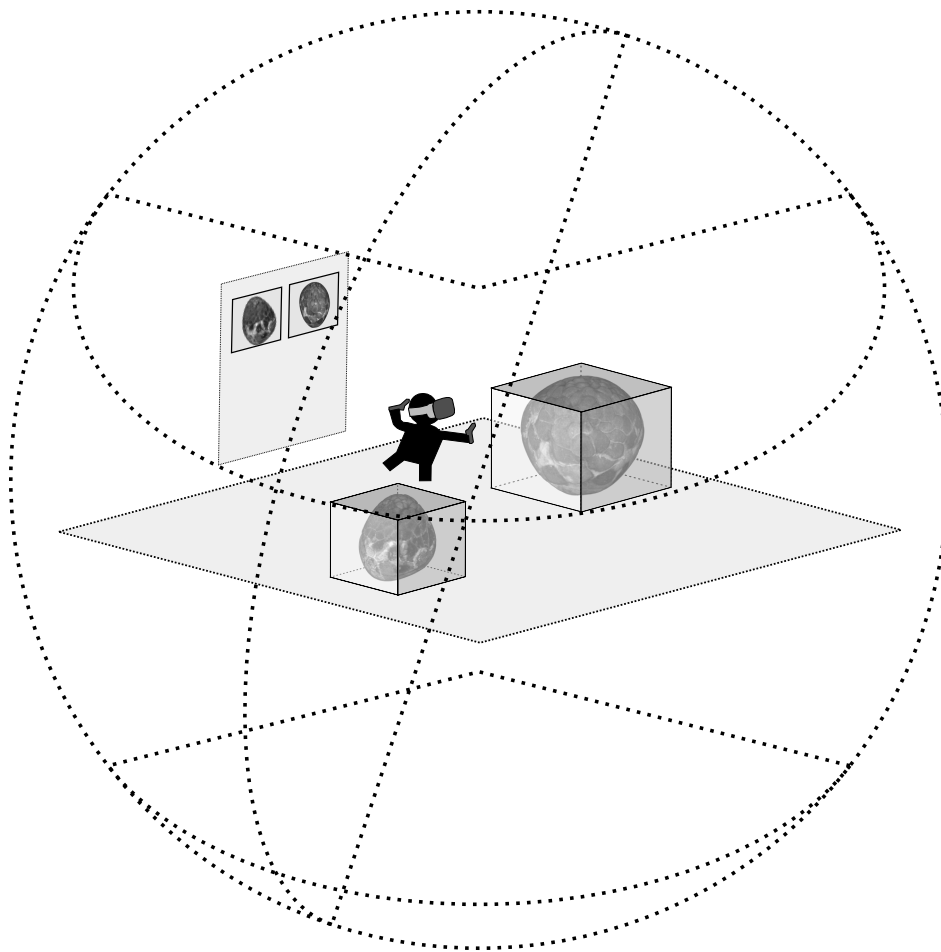
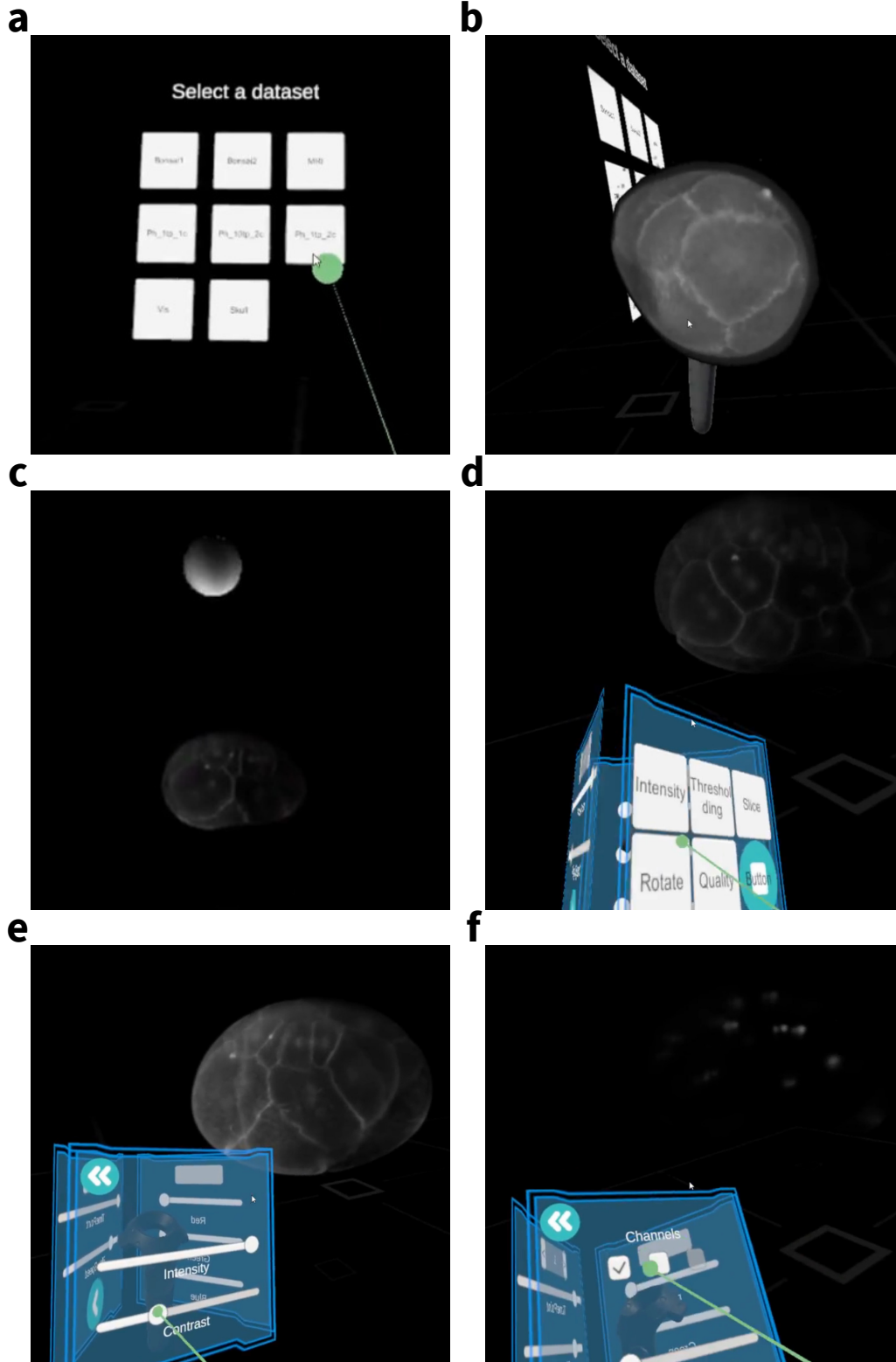


Figure 6.6: Blueprints of the VR room user experience. The user is projected on a transparent platform inside of a dark sphere. Floating is a dataset selection menu allowing the user to open datasets and load them into the room. Each dataset can be placed, scaled, rotated, and normalized at will.

Figure 6.7 (following page): Examples of the VR user experience shown in the form of snapshots of user views from within the VR. a) The dataset selection panel appears on the side of the wall. Each white square displays the name of the dataset defined in the metadata. The user must point the laser towards it to load any dataset. **b)** Example of a volumetric dataset grabbed by the user. When the user selects a dataset on the selection panel, the dataset is first loaded onto the hand of the user. The user can use this to rotate and place the dataset at any location in the room. **c)** Example of a selected dataset. The white hovering sphere indicates that this dataset is now linked to the glove menu. **d)** The hovering glove menu attached to the left arm of the user. The user must point with its right-hand laser onto the buttons to enter any submenu or modify a value. **e)** Example of the user modifying the contrast of the dataset by pointing the laser at the contrast slider and changing the value. The effects are shown in real time thanks to the exposed shader variables. **f)** Example of the channel selector to display multiple channels of the same dataset.

Figure 6.7: (continued)



In a VR experience, the user's hands are visible through controllers, but text input is limited and there is no cursor to select or mouse buttons to click. Moreover, multiple VR experiences had already created intuitive user experience designs that I could use as inspiration. Because it offered a high number of tunable parameters, I decided to build a virtual glove similar to the Google 3D drawing application Tilt Brush⁵⁹. This interface consists of a hovering multipanel interface locked onto the left arm of the user (Figure 6.7). I recreated a similar interface where three panels containing the main user interface (UI) were set to hover on the user's left arm. Using the right controller, a laser pointer would be emitted to serve a cursor to select the different functions. Using the laser pointer, the user was presented with a selection of panels each serving to tune a different set of parameters available in the shader (Figure 6.7). Similarly, the laser pointer was used to select different datasets in the room (Figure 6.7). The selection mechanism worked by creating a hovering sphere above the dataset. This avoided the highlighting of the 3D image, increasing the feeling of immersion. Once a dataset was selected, the menu panel on the virtual glove only affects its parameters.

I performed multiple testing sessions of the interface with three researchers from the Extavour lab as well as over 15 members of the Molecules Cells and Organisms research community during the annual conference of the department in 2018. For each session I first tried to let the user navigate freely, but answered any question they had. I then asked non-standardized questions to collect their feedback on features they had used. This technique is not quantitative therefore no data were collected. Feedback from users (data not shown) was that the virtual glove menu is overall a good design to interact with the parameters. However, it required a high learning curve to understand all the functionalities. Anecdotally, users found the selection cursor to be confusing or not visible enough. Finally, the overall experience of the room was highly appreciated.

6.3.4 Implementing tracking of nuclei in Virtual Reality

The original goal of this project was to not only allow for the intuitive observation of 3D images but also offer a better tracking experience than Mamut⁶⁰. I hypothesized that tracking objects in VR would be more accurate than tracking them in 2D as Mamut allows⁶⁰. As I did not complete

the tracking part of the software, I was not ultimately able to compare the results of tracking in 3D versus 2D. Here I describe the progress that I made towards generating the tracking option of this VR tool.

I first created a visualizer for object tracking annotations by instantiating spheres and cylinders. Each sphere corresponded to a nucleus annotation and each cylinder to a movement from one time point to the next. I based these annotations on the Mamut XML file format⁶⁰. The first version of this visualizer worked well up until 600 annotations, at which point the instantiation of new objects between time points in the virtual environment began to lower the frame rate to below the acceptable VR standard of 90 frames per second (technical challenges of VR frame rates discussed in El Beheiry et al.³⁷).

6.4 Discussion

In this chapter, I presented results towards the creation of a VR software for the observation and annotation of 3D images. VR poses some new challenges in terms of user interfaces and user experiences such as occluding views, locomotion, physics simulations, and object interaction to name a few. One key aspect to successfully develop the user experience was to have repeated testing sessions with different users to gather feedback. A critical aspect of such sessions is to avoid guiding the user so as to try to determine how intuitive the software seems to the user. Though I have not yet collected quantitative data from a user survey, the overall impression is highly positive. One of the main comments offered by users was that the observation of a 3D dataset in a virtual three-dimensional universe gave them a new perspective on the observed biological data. In one instance I offered Extavour lab members the opportunity to look at a gastrulating *P. hawaiiensis* embryo alongside a *G. bimaculatus* gastrulating embryo. The feedback I received from some users was that they could see the difference between invagination (in *G. bimaculatus*) and delamination (in *P. hawaiiensis*) for the first time in a non-schematic representation.

While the capacity to open and observe datasets was completed, many features of this tool remain to be developed. As discussed above, the integration of file formats with the existing central tool BigDataViewer is a key change that must be made to the existing tool in the future,

for users to quickly and easily load their datasets. Moreover, the ability to create annotations and interact with tracked objects is a central feature that should be developed in the future, for this tool to be functional. Therefore, a better coding strategy must be used to achieve the visualization of thousands of annotations. One solution could be to use a particle system, however, particles are non-interactable objects. One key aspect of annotating a dataset is to interact with the annotations by adding, deleting, moving, and scaling them. Particles by default do not allow for such types of interactions. Another solution would be to use object pooling or GPU instantiations. This could allow the user to still interact with the objects and render thousands of them at the same time. Smaller improvements regarding the dynamic loading of datasets should also be implemented such as an automatic out of memory detection and a sliding window of fixed size to dynamically load the samples. The latter will require buffering computations to verify that the playing speed is not faster than buffering time. Finally, continuous user experience changes will need to be implemented such as changes to the selection cursor. Due to the high level of expertise required for the tool a tutorial explaining what each parameter does and how to access them will need to be created. Moreover surveys of customary 3D interaction techniques such as Argelaguet and Andujar⁶¹ can be used to guide the implementation choices.

Finally, in order to show the efficacy of virtual reality, a proper experiment with users will need to be devised. While the subjective perception of volumetric rendering is hard to assess, the tracking of objects in space can be objective. If future work on this project completes the development of 3D tracking features, I suggest that users should be asked users to track nuclei using Mamut. The time it takes the users to perform this tracking should be recorded, and they should be asked to complete a user experience survey with qualitative questions guided by best practice in the field (described in Albert and Tullis⁶²) regarding the experience such as the level of difficulty to learn the tool, or to track objects. Then, the same users should be asked to repeat nucleartracking of the same data in VR. The time it takes them to complete this task should be recorded, and a user feedback survey the same questions should be administered. Whether users are asked to track using Mamut first or the VR tool first, should be randomized to control for prior familiarity with the dataset.

References

- [1] T. Pietzsch, S. Saalfeld, S. Preibisch, and P. Tomancak. BigDataViewer: visualization and processing for large image data sets. *Nat. Methods*, 12(6):481–483, June 2015.
- [2] Blender Foundation. blender.org - Home of the Blender project - Free and Open 3D Creation Software. <https://www.blender.org/>. Accessed: 2020-11-6.
- [3] M. Thompson. The Computer in Art by Jasia Reichardt, and: Cybernetics, Art and Ideas ed. by Jasia Reichardt (review). *Leonardo*, 6(1):81–82, 1973.
- [4] S. Tom. *Moving Innovation: A History of Computer Animation*. MIT Press, 2013.
- [5] A. Silvast and M. Reunanen. Multiple Users, Diverse Users: Appropriation of Personal Computers by Demoscene Hackers. In G. Alberts and R. Oldenziel, editors, *Hacking Europe: From Computer Cultures to Demoscenes*, pages 151–163. Springer London, London, 2014.
- [6] WIRED Staff. Demo or Die! <https://www.wired.com/1995/07/democoders/>, July 1995. Accessed: 2020-10-6.
- [7] D. Blythe. Rise of the Graphics Processor. *Proc. IEEE*, 96(5):761–778, May 2008.
- [8] R. L. Cook. Shade trees. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '84, pages 223–231, New York, NY, USA, January 1984. Association for Computing Machinery.
- [9] P. Hanrahan and J. Lawson. A language for shading and lighting calculations. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '90, pages 289–298, New York, NY, USA, September 1990. Association for Computing Machinery.

- [10] A. A. Apodaca and M. W. Mantle. RenderMan: pursuing the future of graphics. *IEEE Comput. Graph. Appl.*, 10(4):44–49, July 1990.
- [11] C. Maughan and M. Wloka. Vertex shader introduction. *NVIDIA Technical Brief*, 2001.
- [12] C. McClanahan. History and evolution of gpu architecture. *A Survey Paper*, 9:1–7, 2010.
- [13] J. Kessenich, D. Baldwin, and R. Rost. *The opengl shading language Language version 1*. cse.chalmers.se, 2004.
- [14] R. J. Rost. The OpenGL shading language. In *SIGGRAPH Course 17: State-of-the-Art in Hardware Rendering*, pages 6:1–56. SIGGRAPH, 2002.
- [15] I. Quilez and P. Jeremias. Shadertoy. <https://www.shadertoy.com/>, 2013.
- [16] A. Rockwood, K. Heaton, and T. Davis. Real-time rendering of trimmed surfaces. In *Proceedings of the 16th annual conference on Computer graphics and interactive techniques, SIGGRAPH '89*, pages 107–116, New York, NY, USA, July 1989. Association for Computing Machinery.
- [17] K. Akeley and T. Jermoluk. High-performance polygon rendering. In *Proceedings of the 15th annual conference on Computer graphics and interactive techniques, SIGGRAPH '88*, pages 239–246, New York, NY, USA, June 1988. Association for Computing Machinery.
- [18] P. S. Paul, S. Goon, and A. Bhattacharya. History and comparative study of modern game engines. *International Journal of Advanced Computed and Mathematical Sciences*, 3(2): 245–249, 2012.
- [19] M. Toftedahl and H. Engström. A Taxonomy of Game Engines and the Tools that Drive the Industry. In *DiGRA 2019, The 12th Digital Games Research Association Conference, Kyoto, Japan, August, 6-10, 2019*. Digital Games Research Association (DiGRA), 2019.
- [20] M. Toftedahl and H. Engström. A Taxonomy of Game Engines and the Tools that Drive the Industry. In *DiGRA 2019, The 12th Digital Games Research Association Conference, Kyoto, Japan, August, 6-10, 2019*. Digital Games Research Association (DiGRA), 2019.
- [21] T. Sweeney. Unreal Engine, 1995. URL <https://www.unrealengine.com/>.
- [22] D. Helgason, N. Francis, and J. Ante. Unity, 2005. URL <https://unity.com/>.
- [23] J. Linietsky and A. Manzur. Godot Engine, 2007.
- [24] C. Wheatstone. XVIII. Contributions to the physiology of vision. —Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical*

- Transactions of the Royal Society of London*, 128:371–394, January 1838.
- [25] M. L. Heilig. Sensorama simulator. U.S. Patent 3050870, August 1962.
- [26] M. L. Heilig. Stereoscopic-television apparatus for individual use. U.S. Patent 2955156, October 1960.
- [27] I. E. Sutherland. The Ultimate Display. *Proceedings of IFIP Congress*, pages 506–508, 1965.
- [28] I. E. Sutherland. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I, AFIPS '68 (Fall, part I)*, pages 757–764, New York, NY, USA, December 1968. Association for Computing Machinery.
- [29] S. Mandal. Brief introduction of virtual reality & its challenges. *International Journal of Scientific & Engineering Research*, 4(4):304–309, 2013.
- [30] P. Luckey, B. I. Trexler, G. England, and J. McCauley. Virtual reality headset. U.S. Patent D701206:S1, March 2014.
- [31] C. Zellweger, B. E. Barberis, C. S. Kim, I. V. Carl Samuel Conlee, and B. K. N. Robertson. Head mounted display. U.S. Patent D761258:S1, July 2016.
- [32] A. Róžańska. Valve index as a new approach for virtual reality studies. *Zeszyty Naukowe. Elektryka/Politechnika Opolska*, 77:91–94, 2019.
- [33] K. Perlin and E. M. Hoffert. Hypertexture. In *Proceedings of the 16th annual conference on Computer graphics and interactive techniques, SIGGRAPH '89*, pages 253–262, New York, NY, USA, July 1989. Association for Computing Machinery.
- [34] C. Crassin, F. Neyret, S. Lefebvre, and E. Eisemann. GigaVoxels: ray-guided streaming for efficient and detailed voxel rendering. In *Proceedings of the 2009 symposium on Interactive 3D graphics and games, I3D '09*, pages 15–22, New York, NY, USA, February 2009. Association for Computing Machinery.
- [35] J. Beyer, M. Hadwiger, and H. Pfister. State-of-the-art in GPU-based large-scale volume visualization. In *Computer Graphics Forum*, volume 34, pages 13–37, 2015.
- [36] L. H. Williams and T. Drew. What do we know about volumetric medical image interpretation?: a review of the basic science and medical image perception literatures. *Cogn Res Princ Implic*, 4(1):21, July 2019.
- [37] M. El Beheiry, S. Doutreligne, C. Caporal, C. Ostertag, M. Dahan, and J.-B. Masson. Virtual Reality: Beyond Visualization. *J. Mol. Biol.*, 431(7):1315–1321, March 2019.

- [38] W. Feng, H. Zhao, and G. Li. Research on Application of Novel Virtual Reality Technology in Medical Teaching. In *2018 International Conference on Robots Intelligent System (ICRIS)*, volume 1, pages 443–445, May 2018.
- [39] Medical XR. Medical Virtual Reality for Surgery Planning, Medical Education and Patient Information. <https://www.medicalholodeck.com/>, . Accessed: 2020-10-6.
- [40] DICOM VR – Visualizing and Manipulating Medical Imaging in a New Dimension. <http://www.dicomvr.com/>, . Accessed: 2020-10-6.
- [41] Y. Q. Guan, Y. Y. Cai, M. Opas, Z. W. Xiong, and Y. T. Lee. A VR enhanced collaborative system for 3D confocal microscopic image processing and visualization. *Int. J. Image Graph.*, 06(02):231–250, April 2006.
- [42] R. van Liere, W. de Leeuw, J. Mulder, P. Verschure, A. Visser, E. Manders, and R. van Driel. Virtual reality in biological microscopic imaging. In *Proceedings IEEE International Symposium on Biomedical Imaging*, pages 879–882, July 2002.
- [43] Z. Ai, X. Chen, M. Rasmussen, and R. Folberg. Reconstruction and exploration of three-dimensional confocal microscopy data in an immersive virtual environment. *Comput. Med. Imaging Graph.*, 29(5):313–318, July 2005.
- [44] arivis VisionVR. <https://imaging.arivis.com/de/inviewer>, November 2016. Accessed: 2020-11-6.
- [45] B. D. Campbell. Immersive Visualization to Support Scientific Insight. *IEEE Comput. Graph. Appl.*, 36(3):17–21, May 2016.
- [46] R. P. Theart, B. Loos, and T. R. Niesler. Virtual reality assisted microscopy data visualization and colocalization analysis. *BMC Bioinformatics*, 18(S2), 2017.
- [47] Y. Ding, A. Abiri, P. Abiri, S. Li, C.-C. Chang, K. I. Baek, J. J. Hsu, E. Sideris, Y. Li, J. Lee, T. Segura, T. P. Nguyen, A. Bui, R. R. Sevag Packard, P. Fei, and T. K. Hsiai. Integrating light-sheet imaging with virtual reality to recapitulate developmental cardiac mechanics. *JCI Insight*, 2(22), November 2017.
- [48] A. P. R. Johnston, J. Rae, N. Ariotti, B. Bailey, A. Lilja, R. Webb, C. Ferguson, S. Maher, T. P. Davis, R. I. Webb, J. McGhee, and R. G. Parton. Journey to the centre of the cell: Virtual reality immersion into scientific data. *Traffic*, 19(2):105–110, February 2018.
- [49] C. Stefani, A. Lacy-Hulbert, and T. Skillman. ConfocalVR: Immersive Visualization for

- Confocal Microscopy. *J. Mol. Biol.*, 430(21):4028–4035, October 2018.
- [50] A. Spark, A. Kitching, D. Esteban-Ferrer, A. Handa, A. R. Carr, L.-M. Needham, A. Ponjavic, A. M. Santos, J. McColl, C. Leterrier, S. J. Davis, R. Henriques, and S. F. Lee. vLUME: 3D virtual reality for single-molecule localization microscopy. *Nat. Methods*, 17(11):1097–1099, November 2020.
- [51] J. F. Zhang, A. R. Paciorkowski, P. A. Craig, and F. Cui. BioVR: a platform for virtual reality assisted biological data integration and visualization. *BMC Bioinformatics*, 20(1):78, February 2019.
- [52] M. O’Connor, H. M. Deeks, E. Dawn, O. Metatla, A. Roudaut, M. Sutton, L. M. Thomas, B. R. Glowacki, R. Sage, P. Tew, M. Wonnacott, P. Bates, A. J. Mulholland, and D. R. Glowacki. Sampling molecular conformations and dynamics in a multiuser virtual reality framework. *Sci Adv*, 4(6):eaat2731, June 2018.
- [53] W. Usher, P. Klacansky, F. Federer, P.-T. Bremer, A. Knoll, J. Yarch, A. Angelucci, and V. Pascucci. A Virtual Reality Visualization Tool for Neuron Tracing. *IEEE Trans. Vis. Comput. Graph.*, 24(1):994–1003, January 2018.
- [54] Unity Technologies. Unity - Unity. <https://unity.com/frontpage>. Accessed: 2020-10-6.
- [55] VRTK - Virtual Reality Toolkit. <https://www.vrta.io/>, . Accessed: 2020-10-6.
- [56] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona. Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, 9(7):676–682, July 2012.
- [57] J. Krüger, K. Potter, R. S. Macleod, and C. Johnson. Uvf - Unified Volume Format: A General System for Efficient Handling of Large Volumetric Datasets. *IEEE Conf. Inf. Vis.*, 2008:19–26, 2008.
- [58] C. Sommer, C. Straehle, U. Köthe, and F. A. Hamprecht. Ilastik: Interactive learning and segmentation toolkit. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 230–233, March 2011.
- [59] Tilt Brush. <https://www.tiltbrush.com/>, . Accessed: 2020-10-6.
- [60] C. Wolff, J.-Y. Tinevez, T. Pietzsch, E. Stamatakis, B. Harich, L. Guignard, S. Preibisch,

- S. Shorte, P. J. Keller, P. Tomancak, and A. Pavlopoulos. Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. *Elife*, 7:e34410, March 2018.
- [61] F. Argelaguet and C. Andujar. A survey of 3D object selection techniques for virtual environments. *Comput. Graph.*, 37(3):121–136, May 2013.
- [62] W. Albert and T. Tullis. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Newnes, May 2013.

*We shall not cease from exploration, and the end of all
our exploring will be to arrive where we started and
know the place for the first time.*

Thomas Stearns Eliot

7

Conclusion

7.1 The era of omics

I conducted my doctoral research soon after a technological revolution¹ which led to a radical shift in how researchers generated data about biological systems (reviewed by Heather and Chain²). The creation of technologies to sequence organisms' genomes and transcriptomes led to the generation of large sequence databases (such as the Sequence Read Archive described in Kodama et al.³). In the span that it took from the first analysis to the preparation of the manuscript for Chapter 2, the number of available insect transcriptomes nearly doubled (See Chapter 2 Methods: *Genome and transcriptome preprocessing*). Not only do we now have access to sequence information from many more species, but the nature of those datasets has also changed, with the resolution at which we measure gene expression now reaching the single cell^{4,5,6,7,8}. Furthermore, the continuously accumulating knowledge of decades of research on

the regulatory information of genes and protein-protein interactions has now been synthesized into databases such as FlyBase⁹, KEGG¹⁰ or BioGrid¹¹. The large amount of information available created new avenues for complex system approaches to biology^{12,13,14,15,16} and new microscopy techniques transformed the imaging of biological samples^{17,18,19,20,21,22}. Finally, the way we can see, interpret, and analyze high dimensionality data is about to be transformed by new visualization technologies such as Virtual Reality (See Chapter 6). When reality itself can be molded to be an expression of creative visualization, I believe that the possibilities to understand, teach, and intuit novel ideas increase drastically. In this era of omics, I composed my Ph.D. around the analysis of previously published datasets, the remapping of large-scale novel phenotypic data onto existing datasets, and the generation of new "omics" datasets.

7.2 *oskar*, a novel gene with an intriguing evolution

In Chapter 1 we showed that the origin of *oskar* resulted from a horizontal gene transfer followed by a fusion with a eukaryotic domain²³. To my knowledge, this is only the second time such a mechanism for the formation of a new gene was described²⁴. This study was made possible due to the increase in sequence diversity in databases allowing for a more accurate reconstruction of the evolutionary history of genes. Two key aspects of Chapter 1 were aided by this diversity. First, it allowed me to collect over 100 *oskar* ortholog sequences which brought statistical power to the amino acid sequence information of both the LOTUS and OSK domains. One of the key techniques I used throughout Chapter 1 and Chapter 2 was the modeling of sequence data by a Hidden Markov Model^{25,26}. This model becomes more accurate with the increasing number and diversity of sequence information used to create it. The use of this model was crucial to find bacterial genes with likely homology to *oskar* and especially to the OSK domain. Second, the large collection of sequence data from all kingdoms of life made the collection of sequences and the phylogenetic analysis more robust. It is always difficult to assess whether an absence of homologs is due to effects like sequence divergence and taxonomic undersampling. With increased sampling, the correlation between an absence of hits and the real absence of homologs increases with each sequence added. This also affected the phylogenetic reconstruction of LOTUS and OSK, and statistical tools such as SOWHAT²⁷. Indeed,

both use maximum likelihood estimation^{27,28} or Bayesian statistics²⁹ which compute estimated distributions from the underlying alignment, therefore are sensitive to sample size. Finally, recent studies are starting to show the importance and prevalence of Horizontal Gene Transfers^{30,31,32,33,34}. By proposing this new mechanism, I hope that further studies will look for similar evolutionary events for new genes of unknown origins.

Not only did *oskar* turn out to have an intriguing origin, but its evolutionary history was mostly unknown^{23,35}. Therefore in Chapter 2, we studied the sequence evolution of *oskar*. We proposed hypotheses as to the putative functions of conserved motifs, which are testable and could be the subject of future studies including biochemical and mutational experiments. One of the main questions behind Chapter 2 was whether we could observe differences between Oskar sequences from the hemimetabolous insects and holometabolous insects. While the LOTUS domain did not show any differential conservation, the OSK domain was more conserved in hemimetabolous sequences. Given the importance of *oskar* in recruiting key mRNAs to the posterior pole of the embryo³⁶ through the binding of mRNA via the RNA binding of the OSK domain³⁷, I would have expected this result to be reversed. However, to form a liquid-liquid phase separation such as germ plasm, it was described that as the valency between two interacting proteins needed to reach a critical threshold³⁸. My hypothesis is that the need to increase valency to form germ plasm³⁸ led to a relaxation of the selective pressure on OSK (as measured by the lower overall conservation) to allow for more RNA partners. These results should be further studied by expanding the biochemical analysis of hemimetabolous Oskar. For example, future studies should identify the primary partners of Oskar in the neuroblasts of developing *Gryllus bimaculatus* embryos. With respect to reported observations of OSK's ability to interact with the 3'UTR of specific mRNAs³⁷, future studies should determine whether the rate of evolution of the 3'UTR of *oskar* is particularly high when compared to other 3'UTRs in *D. melanogaster*, and whether there is any significant degree of co-evolution between OSK and the 3'UTRs of *oskar* partners.

Another key biochemical property of *D. melanogaster* Oskar is the dimerization of the LOTUS domain³⁹. We found that while the amino acid sequence of this domain was not highly conserved at the dimerization interface, the LOTUS domains from insect Oskars that were predicted to dimerize³⁹ showed high physicochemical conservation compared to the

monomeric ones. In order to follow up on this result, a machine learning approach using a classifier could be trained on the predicted dimerizing and monomeric sequences and then used to predict the dimerization properties of the other Oskar LOTUS sequences. To confirm the prediction of the classifier, in vitro expression of soluble full length Oskar protein (or at a minimum expression of the LOTUS domain) and biochemical confirmation of the dimerization via SDS-PAGE or size elution chromatography^{39,40} would be needed. Furthermore, we also found that the interface with Vasa showed high sequence conservation across all insects. To further explore the evolution of this interface, a sequence co-evolution analysis between the exposed amino acids of LOTUS and Vasa could shed light on the exact biochemical nature of their binding^{41,42}.

In both Chapter 1 and Chapter 2, the automation process was key for the discovery and analysis of *oskar* orthologs. While sequence databases of orthologous gene groups such as EggNog⁴³ and OrthoDB⁴⁴ provide invaluable information, they rely on well-defined Hidden Markov Model representations of sequence groups. They also rely on the correct annotation of genomes and assembly of transcriptomes. In the case of new genes or less-studied genes, such an automated process might fail to define an orthologous group. This is the case with *oskar*, for example: in OrthoDB only 51 (search performed on the 10/25/2020) sequences are described. The careful building of a gene model for *oskar* that I carried out in Chapter 1 allowed us to automate the discovery process for Chapter 2 and was the only necessarily manual step of the process. I hope that further studies will continue to automate such processes for less-studied genes.

7.3 The regulation of the development of the *Drosophila* ovary

When Extavour lab postdoc Tarun Kumar presented the first results of the screen he was performing on *D. melanogaster*, we decided to collaborate together. The amount of phenotypic data he was collecting allowed us to conduct a systematic analysis of the effect of signaling pathways on the regulation of ovariole number and egg-laying capacity, which constituted the work presented in Chapter 3. Our discovery of putative gene modules underlying the regulation

of ovariole numbers and egg laying is a stepping stone for further studies on the genetic control of both traits. In the future, combining the ovariole number information collected from Hawaiian *Drosophila*⁴⁵ with genome or transcriptome data could make it possible to determine whether the evolution of network topology is linked to the evolution of ovariole number. By hypothesizing that the protein-protein interaction network is conserved throughout drosophilids, network attacks⁴⁶ (a process by which nodes are removed from a network and the topological integrity and resilience is measured) can be performed.

Another conclusion that we drew from this study was that all described animal signaling pathways play a role in the regulation of both phenotypes. In my opinion, this is not a surprising result as I expected that the formation of an organ such as the ovary would require the interplay of many factors. I believe that future studies should apply similar approaches to other phenotypes, and I expect that similar results showing a complex web of interconnected regulatory mechanisms will emerge.

Finally, previous works have successfully shown that dynamical modeling can be achieved with only a directed network topology⁴⁷. One of the promising new avenues this study generated is the creation of dynamical regulatory models by using the known interactions between signaling pathway genes. Using the underlying network topology found in our study, I predict that dynamical models of ovariole number regulation can be acquired by fitting the models published in Santolini and Barabási⁴⁷ onto the measured phenotypic data. However, those models were generated and tested on single cells only, therefore their extension to the regulation of tissues will likely require multiple adjustments.

7.4 Towards a better understanding of germ layer specification in *P. hawaiiensis*

When I started my Ph.D., I proposed an ambitious plan in which I would generate single-cell libraries for multiple time points of the first three days of wild type *P. hawaiiensis* development, image live embryos in 4D the same period of time, and merge both datasets to create an interactive virtual 4D gene expression atlas. In Chapter 4, I discussed the advances made

towards obtaining single-cell transcriptomic data of developing embryos, but did not achieve the goals originally set for Chapter 4. Despite the difficulties encountered, the encapsulation and sequencing of a small number of cells was detected (as per the filtering and selection for cells in the library explained in Chapter 4). I would recommend collaboration with a chemist to determine new enzymatic or chemical methods to digest the egg shell. Moreover, a high quality embryonic transcriptome should be generated before attempting this experiment again in the future, to alleviate the difficulties that I encountered during genome annotation. If a transcriptome were generated at multiple embryonic time points, it could also be used to explore the gene expression changes happening during the first three days of embryogenesis. Depending on the observed expression changes in these whole-embryo transcriptomes over time, the definition of the number of time points needed for subsequent single cell RNA sequencing could be further refined. To conclude, while I did not generate libraries of sufficient quality for the questions I set out to answer, I demonstrated the in-principle feasibility of this approach, and identified important technical bottlenecks that would have to be overcome by anyone wishing to carry these experiments forward. I hope that future work will continue to expand on this process and fully succeed at generating libraries.

In Chapter 5, I recorded in 4D the first three and four days of four wild type developing *P. hawaiiensis* embryos. The high spatio-temporal resolution of light sheet microscopy allowed me to successfully track one of the four embryos, generating a ground truth dataset for subsequent analysis. The generation of those tracks was more time consuming than I had planned for. At the time I was doing this analysis, to my knowledge there were few options to automatically segment and track embryos in three dimensions. I used the Ilastik random forest classifier⁴⁸ on the first embryo, but it yielded unusable tracks that required me to check the entire embryo, such that I believe it would likely have been faster to track it manually. The other package that I was aware of at the time was TGMM, a gaussian kernel segmentation and tracking algorithm tailored for light sheet datasets⁴⁹. However, despite my best efforts, I found it impossible to compile this package.

At the time of writing this thesis, two new packages using an artificial neural network architecture were published, StarDist and CellPose^{50,51}. StarDist⁵⁰ uses a U-Net and ResNet architecture as a first segmentation step and to map a cell border probability, so I expect that

good results could be achieved with this new algorithm if it were used to analyze the *P. hawaiiensis* lightsheet data that I collected. Indeed, during my Ph.D. as an exercise in learning artificial neural networks I trained U-Net architecture⁵² on a small ground truth dataset which resulted in a 97% pixel classification accuracy (data not shown). However, I could not use the native U-Net architecture as it is designed for 2D images and does not allow for the subsequent separation of objects. Both CellPose and StarDist have added subsequent layers to the segmentation to allow for the separation of single objects^{50,51}. Thanks to the fully tracked embryo that I was able to analyze, it should, in principle, be possible to extract an appropriately large amount of data to train those networks. Therefore I believe that the segmentation of the nuclei will not be a bottleneck issue in the near future. Finally, even if the 3D positions of the cells are known at a given time point, their tracking in 4D is a current challenge⁵³. The current top performing algorithm on lightsheet datasets according to the Cell Tracking Benchmark⁵³ is the Baxter Algorithm, which defines probability functions for events like migration, division, or death and maximizes a scoring function (akin to information entropy maximization)⁵⁴. However, the more recent TGMM2 and Vector Flow algorithm⁵⁵ were not part of the challenge, therefore it is hard to know how the Baxter algorithm compares to them on light sheet data. In the second part of Chapter 5, I described an ablation experiment aimed at understanding the intra-germ layer compensation mechanism previously reported for *P. hawaiiensis*⁵⁶. To analyse the resulting data, in addition to the segmentation and tracking challenges described above for wild type embryos, a custom surface reprojection will be needed to compare the ablated embryos with each other and with wild type embryos. *P. hawaiiensis* embryos are roughly spherical, which helps with morphometric analysis (see Chapter 5 Methods: Mercator projection of nuclei onto a 2D space). Ablated embryos, however, do not recuperate their spherical shape within the three days that we performed our imaging. To compare the wild type and ablated embryos, a custom surface will therefore need to be created to project the position of the cells onto a 2D or 3D map. Software like IMSane should allow for such a reprojection in 2D⁵⁷. Another possible method would be to use a 3D lattice transformer, as was used in the comparison of developing mouse embryos⁵⁵. Due to the invariant early cleavage nature of *P. hawaiiensis* embryos, understanding the morphogenetic changes that compensate for the lack of a blastomere will increase our understanding of the establishment of cellular territories and

body plans in this crustacean.

To conclude, while I made progress towards the understanding of the germ layer specification of *P. hawaiiensis*, I could not complete the project. Nonetheless, I generated valuable datasets that can be further analyzed, and could help to provide answers to those questions in the future.

7.5 Expanding our perception with a third dimension

In the final chapter, Chapter 6, of this thesis, I took a detour from biology into the field of computer graphics and virtual reality. This was motivated by the absence of open visualization interfaces to observe and annotate in 3D the light sheet dataset generated in Chapter 5. The first part of this tool was completed successfully, namely a tool to permit the observation and manipulation of 3D images in a virtual space. User feedback was uniformly positive (see Chapter 6), and the software source was published in a Open Source licence (https://gitlab.com/xqua/microscopy_vr), allowing for anyone to extend its base. This tool also allows for users to import and render tracks generated with Mamut¹⁷. However, the necessary features to track cells within the tool were not completed yet, because I found that creating a custom tracking mechanism in VR was a more complex task than the visualization of tracks.

In the future, upon completion of this project, I plan to perform a user survey comparing user perception and speed of annotation between state of the art tracking software BigDataViewer⁵⁸ and Mamut¹⁷ and my new VR tool, to quantify the difference. I am convinced that 3D tracking will be faster than 2D tracking. I plan to continue the development of this software after my Ph.D., to be able to release a more complete version with features such as importing datasets from BigDataViewer⁵⁸ and allowing multiple users to interact with a dataset in real time. To conclude, while this was a detour from the main biological questions of my thesis work, it was a very rewarding project that I believe will serve the biological community.

References

- [1] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.*, 55(4):641–658, April 2009.
- [2] J. M. Heather and B. Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, January 2016.
- [3] Y. Kodama, M. Shumway, R. Leinonen, and International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, 40(Database issue):D54–6, January 2012.
- [4] T. Hashimshony, F. Wagner, N. Sher, and I. Yanai. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, 2(3):666–673, September 2012.
- [5] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.
- [6] T. Hashimshony, N. Senderovich, G. Avital, A. Klochender, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, Y. Dor, A. Regev, and I. Yanai. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, 17:77, April 2016.
- [7] E. Z. Macosko, A. Basu, R. Satija, J. Nemes, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015.
- [8] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D.

- Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, January 2017.
- [9] J. Thurmond, J. L. Goodman, V. B. Strelets, H. Attrill, L. S. Gramates, S. J. Marygold, B. B. Matthews, G. Millburn, G. Antonazzo, V. Trovisco, T. C. Kaufman, B. R. Calvi, and FlyBase Consortium. FlyBase 2.0: the next generation. *Nucleic Acids Res.*, 47(D1):D759–D765, January 2019.
- [10] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, 44(D1):D457–62, January 2016.
- [11] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34(Database issue):D535–9, January 2006.
- [12] R.-S. Wang, K. T. Hall, F. Giulianini, D. Passow, T. J. Kaptchuk, and J. Loscalzo. Network analysis of the genomic basis of the placebo effect. *JCI Insight*, 2(11):93911, June 2017.
- [13] S. D. Ghiassian, J. Menche, and A.-L. Barabási. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.*, 11(4):e1004120, April 2015.
- [14] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, April 2004.
- [15] E. Guney, J. Menche, M. Vidal, and A.-L. Barabási. Network-based in silico drug efficacy screening. *Nat. Commun.*, 7:10331, February 2016.
- [16] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.
- [17] C. Wolff, J.-Y. Tinevez, T. Pietzsch, E. Stamatakis, B. Harich, L. Guignard, S. Preibisch, S. Shorte, P. J. Keller, P. Tomancak, and A. Pavlopoulos. Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod

- limb. *Elife*, 7:e34410, March 2018.
- [18] R. K. Chhetri, F. Amat, Y. Wan, B. Höckendorf, W. C. Lemon, and P. J. Keller. Whole-animal functional and developmental imaging with isotropic spatial resolution. *Nat. Methods*, 12(12):1171–1178, December 2015.
- [19] P. J. Keller, A. D. Schmidt, A. Santella, K. Khairy, Z. Bao, J. Wittbrodt, and E. H. K. Stelzer. Fast, high-contrast imaging of animal development with scanned light sheet-based structured-illumination microscopy. *Nat. Methods*, 7(8):637–642, August 2010.
- [20] M. B. Ahrens, M. B. Orger, D. N. Robson, J. M. Li, and P. J. Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods*, 10(5):413–420, May 2013.
- [21] B.-C. Chen, W. R. Legant, K. Wang, L. Shao, D. E. Milkie, M. W. Davidson, C. Janetopoulos, X. S. Wu, J. A. Hammer, 3rd, Z. Liu, B. P. English, Y. Mimori-Kiyosue, D. P. Romero, A. T. Ritter, J. Lippincott-Schwartz, L. Fritz-Laylin, R. D. Mullins, D. M. Mitchell, J. N. Bembenek, A.-C. Reymann, R. Böhme, S. W. Grill, J. T. Wang, G. Seydoux, U. S. Tulu, D. P. Kiehart, and E. Betzig. Lattice light-sheet microscopy: imaging molecules to embryos at high spatiotemporal resolution. *Science*, 346(6208):1257998, October 2014.
- [22] A. W. Bisson-Filho, Y.-P. Hsu, G. R. Squyres, E. Kuru, F. Wu, C. Jukes, Y. Sun, C. Dekker, S. Holden, M. S. VanNieuwenhze, Y. V. Brun, and E. C. Garner. Treadmilling by FtsZ filaments drives peptidoglycan synthesis and bacterial cell division. *Science*, 355(6326):739–743, February 2017.
- [23] L. Blondel, T. E. Jones, and C. G. Extavour. Bacterial contribution to genesis of the novel germ line determinant *oskar*. *Elife*, 9:e45539, February 2020.
- [24] J. Schultz. HTTM, a horizontally transferred transmembrane domain. *Trends Biochem. Sci.*, 29(1):4–7, January 2004.
- [25] S. R. Eddy. Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 3:114–120, 1995.
- [26] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.
- [27] S. H. Church, J. F. Ryan, and C. W. Dunn. Automation and Evaluation of the SOWH Test with SOWHAT. *Syst. Biol.*, 64(6):1048–1058, November 2015.
- [28] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large

- phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.
- [29] J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, August 2001.
- [30] J. C. Dunning Hotopp, M. E. Clark, D. C. S. G. Oliveira, J. M. Foster, P. Fischer, M. C. Muñoz Torres, J. D. Giebel, N. Kumar, N. Ishmael, S. Wang, J. Ingram, R. V. Nene, J. Shepard, J. Tomkins, S. Richards, D. J. Spiro, E. Ghedin, B. E. Slatko, H. Tettelin, and J. H. Werren. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, 317(5845):1753–1756, September 2007.
- [31] P. J. Keeling and J. D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.*, 9(8):605–618, August 2008.
- [32] A. C. C. Wilson and R. P. Duncan. Signatures of host/symbiont genome coevolution in insect nutritional endosymbioses. *Proceedings of the National Academy of Sciences*, 112(33):10255–10261, 2015.
- [33] F. Husnik, N. Nikoh, R. Koga, L. Ross, R. P. Duncan, M. Fujie, M. Tanaka, N. Satoh, D. Bachtrog, A. C. C. Wilson, C. D. von Dohlen, T. Fukatsu, and J. P. McCutcheon. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, 153(7):1567–1578, June 2013.
- [34] D. B. Sloan, A. Nakabachi, S. Richards, J. Qu, S. C. Murali, R. A. Gibbs, and N. A. Moran. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Mol. Biol. Evol.*, 31(4):857–871, April 2014.
- [35] J. A. Lynch, O. Ozüak, A. Khila, E. Abouheif, C. Desplan, and S. Roth. The phylogenetic origin of *oskar* coincided with the origin of maternally provisioned germ plasm and pole cells at the base of the Holometabola. *PLoS Genet.*, 7(4):e1002029, April 2011.
- [36] A. Ephrussi, L. K. Dickinson, and R. Lehmann. Oskar organizes the germ plasm and directs localization of the posterior determinant nanos. *Cell*, 66(1):37–50, July 1991.
- [37] N. Yang, Z. Yu, M. Hu, M. Wang, R. Lehmann, and R.-M. Xu. Structure of *Drosophila* Oskar reveals a novel RNA binding protein. *Proc. Natl. Acad. Sci. U. S. A.*, 112(37):11541–11546, September 2015.
- [38] P. Li, S. Banjade, H.-C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q.-X. Jiang, B. T. Nixon, and M. K. Rosen. Phase

- transitions in the assembly of multivalent signalling proteins. *Nature*, 483(7389):336–340, March 2012.
- [39] M. Jeske, M. Bordi, S. Glatt, S. Müller, V. Rybin, C. W. Müller, and A. Ephrussi. The Crystal Structure of the *Drosophila* Germline Inducer Oskar Identifies Two Domains with Distinct Vasa Helicase- and RNA-Binding Activities. *Cell Reports*, 12(4):587–598, 2015.
- [40] M. Jeske, C. W. Müller, and A. Ephrussi. The LOTUS domain is a conserved DEAD-box RNA helicase regulator essential for the recruitment of Vasa to the germ plasm and nuage. *Genes Dev.*, 31(9):939–952, May 2017.
- [41] T. A. Hopf, C. P. I. Schärfe, J. P. G. L. M. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. J. J. Bonvin, and D. S. Marks. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, 3:e03430, September 2014.
- [42] S. de Oliveira and C. Deane. Co-evolution techniques are reshaping the way we do structural bioinformatics. *F1000Res.*, 6:1224, July 2017.
- [43] J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering, and P. Bork. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, 47(D1):D309–D314, November 2018.
- [44] E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. Simão, and E. M. Zdobnov. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.*, 47(D1):D807–D811, November 2018.
- [45] D. P. Sarikaya, S. H. Church, L. P. Lagomarsino, K. N. Magnacca, S. L. Montgomery, D. K. Price, K. Y. Kaneshiro, and C. G. Extavour. Reproductive Capacity Evolves in Response to Ecology through Common Changes in Cell Number in Hawaiian *Drosophila*. *Curr. Biol.*, 29(11):1877–1884.e6, June 2019.
- [46] A. Kılıç, M. Santolini, T. Nakano, M. Schiller, M. Teranishi, P. Gellert, Y. Ponomareva, T. Braun, S. Uchida, S. T. Weiss, A. Sharma, and H. Renz. A systems immunology approach identifies the collective impact of 5 miRs in Th2 inflammation. *JCI Insight*, 3(11), June 2018.
- [47] M. Santolini and A.-L. Barabási. Predicting perturbation patterns from the topology of

- biological networks. *Proc. Natl. Acad. Sci. U. S. A.*, 115(27):E6375–E6383, July 2018.
- [48] C. Sommer, C. Straehle, U. Köthe, and F. A. Hamprecht. Ilastik: Interactive learning and segmentation toolkit. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 230–233, March 2011.
- [49] F. Amat, W. Lemon, D. P. Mossing, K. McDole, Y. Wan, K. Branson, E. W. Myers, and P. J. Keller. Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nat. Methods*, 11(9):951–958, September 2014.
- [50] M. Weigert, U. Schmidt, R. Haase, K. Sugawara, and G. Myers. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3666–3673, 2020.
- [51] C. Stringer, M. Michaelos, and M. Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *bioRxiv*, page 931238, February 2020.
- [52] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*, cs.CV:1505.04597, May 2015.
- [53] V. Ulman, M. Maška, K. E. G. Magnusson, O. Ronneberger, C. Haubold, N. Harder, P. Matula, P. Matula, D. Svoboda, M. Radojevic, I. Smal, K. Rohr, J. Jaldén, H. M. Blau, O. Dzyubachyk, B. Lelieveldt, P. Xiao, Y. Li, S.-Y. Cho, A. C. Dufour, J.-C. Olivo-Marin, C. C. Reyes-Aldasoro, J. A. Solis-Lemus, R. Bensch, T. Brox, J. Stegmaier, R. Mikut, S. Wolf, F. A. Hamprecht, T. Esteves, P. Quelhas, Ö. Demirel, L. Malmström, F. Jug, P. Tomancak, E. Meijering, A. Muñoz-Barrutia, M. Kozubek, and C. Ortiz-de Solorzano. An objective comparison of cell-tracking algorithms. *Nat. Methods*, 14(12):1141–1152, December 2017.
- [54] K. E. G. Magnusson, J. Jaldén, P. M. Gilbert, and H. M. Blau. Global linking of cell tracks using the Viterbi algorithm. *IEEE Trans. Med. Imaging*, 34(4):911–929, April 2015.
- [55] K. McDole, L. Guignard, F. Amat, A. Berger, G. Malandain, L. A. Royer, S. C. Turaga, K. Branson, and P. J. Keller. In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. *Cell*, 175(3):859–876.e33, October 2018.
- [56] A. L. Price, M. S. Modrell, R. L. Hannibal, and N. H. Patel. Mesoderm and ectoderm lineages in the crustacean *Parhyale hawaiiensis* display intra-germ layer compensation. *Dev. Biol.*, 341(1):256–266, May 2010.
- [57] I. Heemskerk and S. J. Streichan. Tissue cartography: compressing bio-image data by

dimensional reduction. *Nat. Methods*, 12(12):1139–1142, December 2015.

- [58] T. Pietzsch, S. Saalfeld, S. Preibisch, and P. Tomancak. BigDataViewer: visualization and processing for large image data sets. *Nat. Methods*, 12(6):481–483, June 2015.



Chapter 1: Supplementary data

This appendix contains the supplementary figures for the Chapter 1.

Figure A.1 (following page): LOTUS Domain RaxML MUSCLE Tree. Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the LOTUS alignment HMM model. The top 97 hits were selected for phylogenetic analysis, and the only three bacterial sequences found to be a match were added to the alignment manually. The resulting 100 sequences were aligned using MUSCLE with default settings. The sequences were filtered to contain only one sequence per species (best E-value kept) yielding 100 sequences for analysis. Finally, the tree was created using RaxML v8.2.4, using 1000 bootstraps and model selection performed by the RaxML automatic model selection tool. See Section 1.4.5 **Phylogenetic Analysis Based on MUSCLE Alignment** for further detail. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure A.1: (continued)

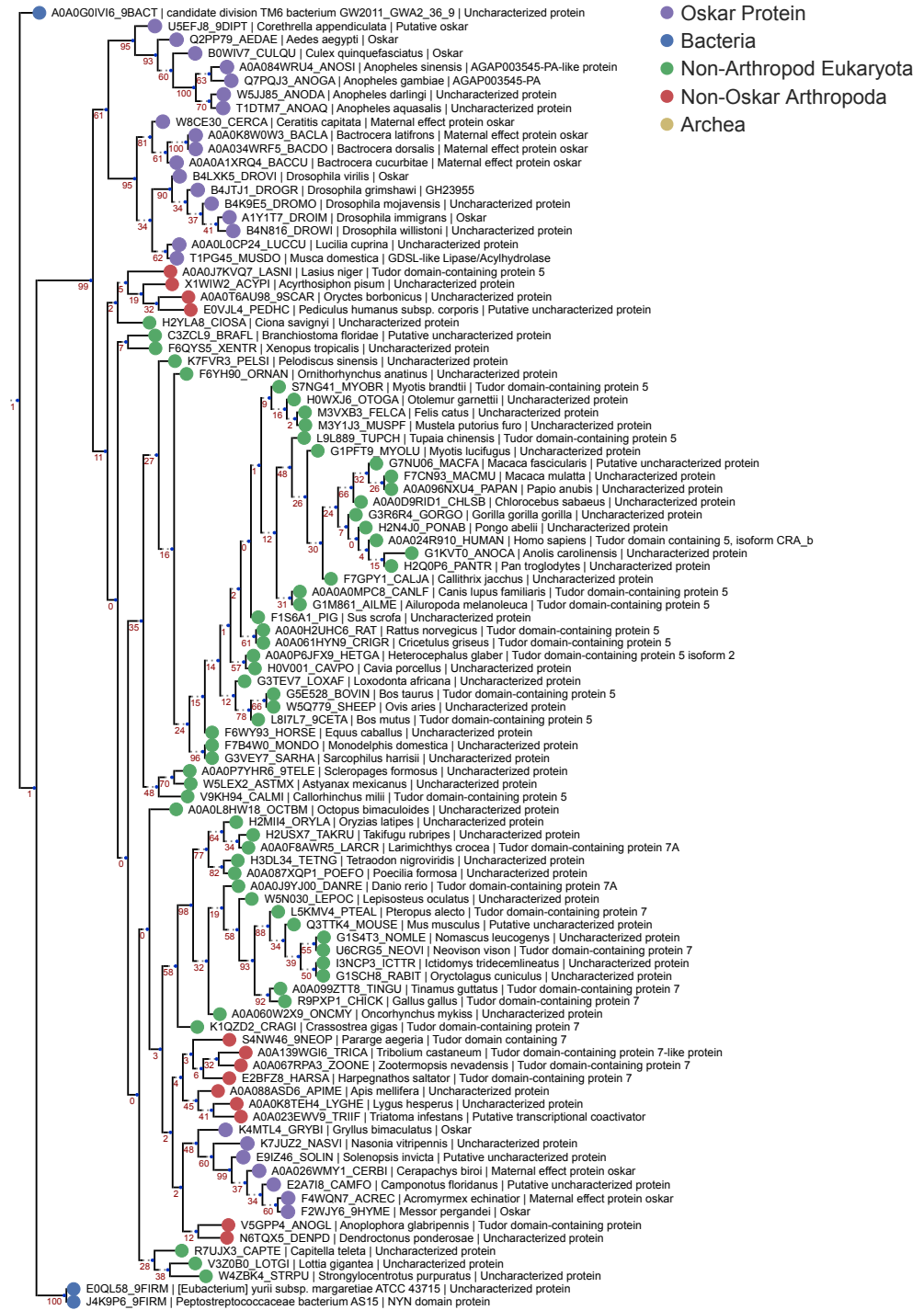


Figure A.2 (following page): LOTUS Domain Bayesian MUSCLE Tree. Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the LOTUS alignment HMM model. 100 sequences were chosen for analysis as described for Figure A.1. The tree was created using Mr Bayes V3.2.6 using a Mixed model (prset aamodel=Mixed) and a gamma distribution (lset rates=Gamma). The algorithm was allowed to run for 3 million generations to achieve a std < 0.01. See Section 1.4.5 **Phylogenetic Analysis Based on MUSCLE Alignment** for further detail. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure A.2: (continued)

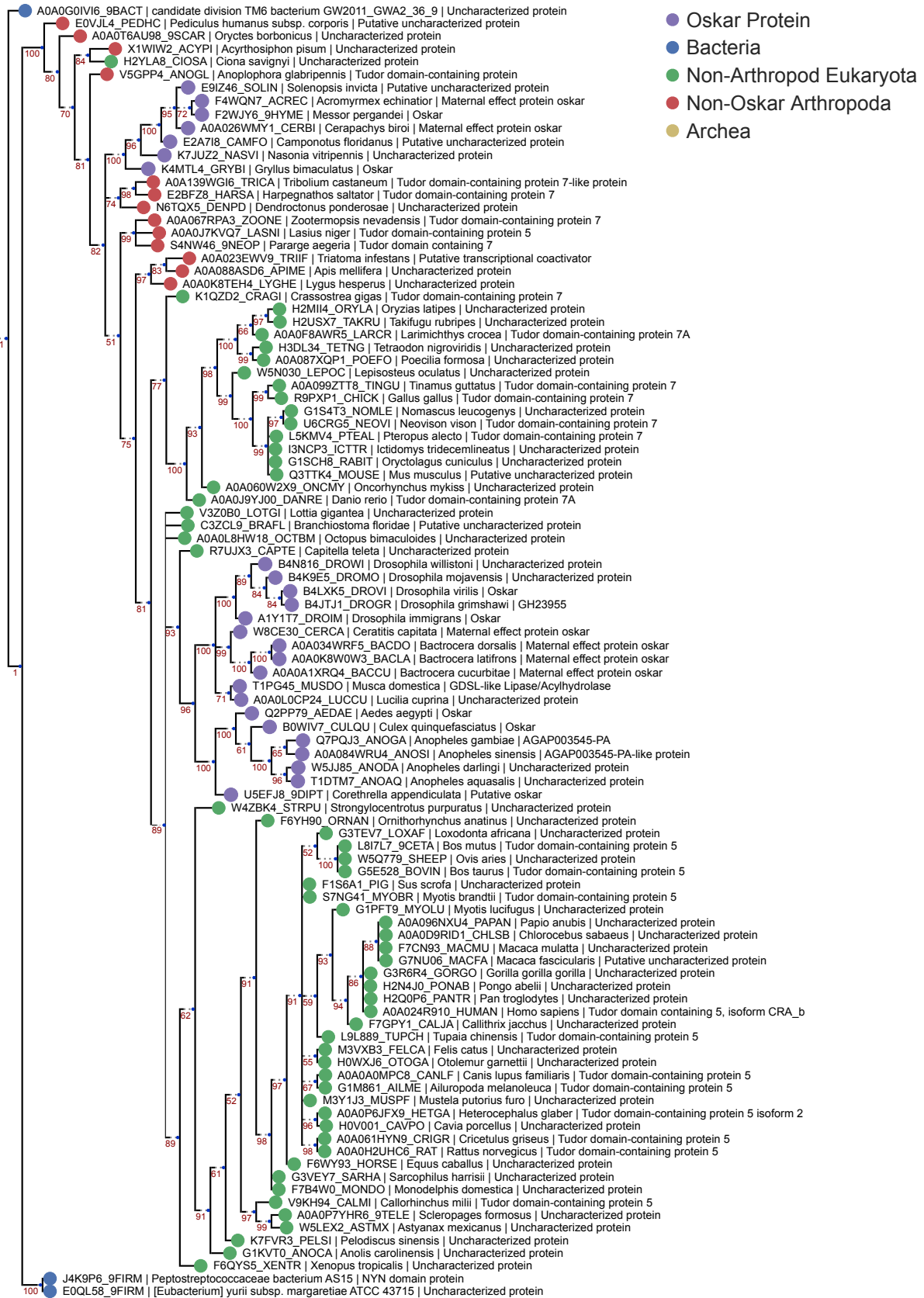


Figure A.3 (following page): OSK Domain RaxML MUSCLE Tree. Phylogenetic tree of the HMMER sequences retrieved from the UniProt database using the OSK alignment HMM model. The top 95 hits were selected for phylogenetic analysis, and the only five non-Oskar eukaryotic sequences found to be a match were added to the alignment manually. The resulting 100 sequences were aligned using MUSCLE with default settings. The sequences were filtered to contain only one sequence per species (best E-value kept), yielding 87 sequences for analysis. Finally, the tree was created using RaxML v8.2.4, using 1000 bootstraps and model selection performed by the RaxML automatic model selection tool. See Section 1.4.5 **Phylogenetic Analysis Based on MUSCLE Alignment** for further detail. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure A.3: (continued)

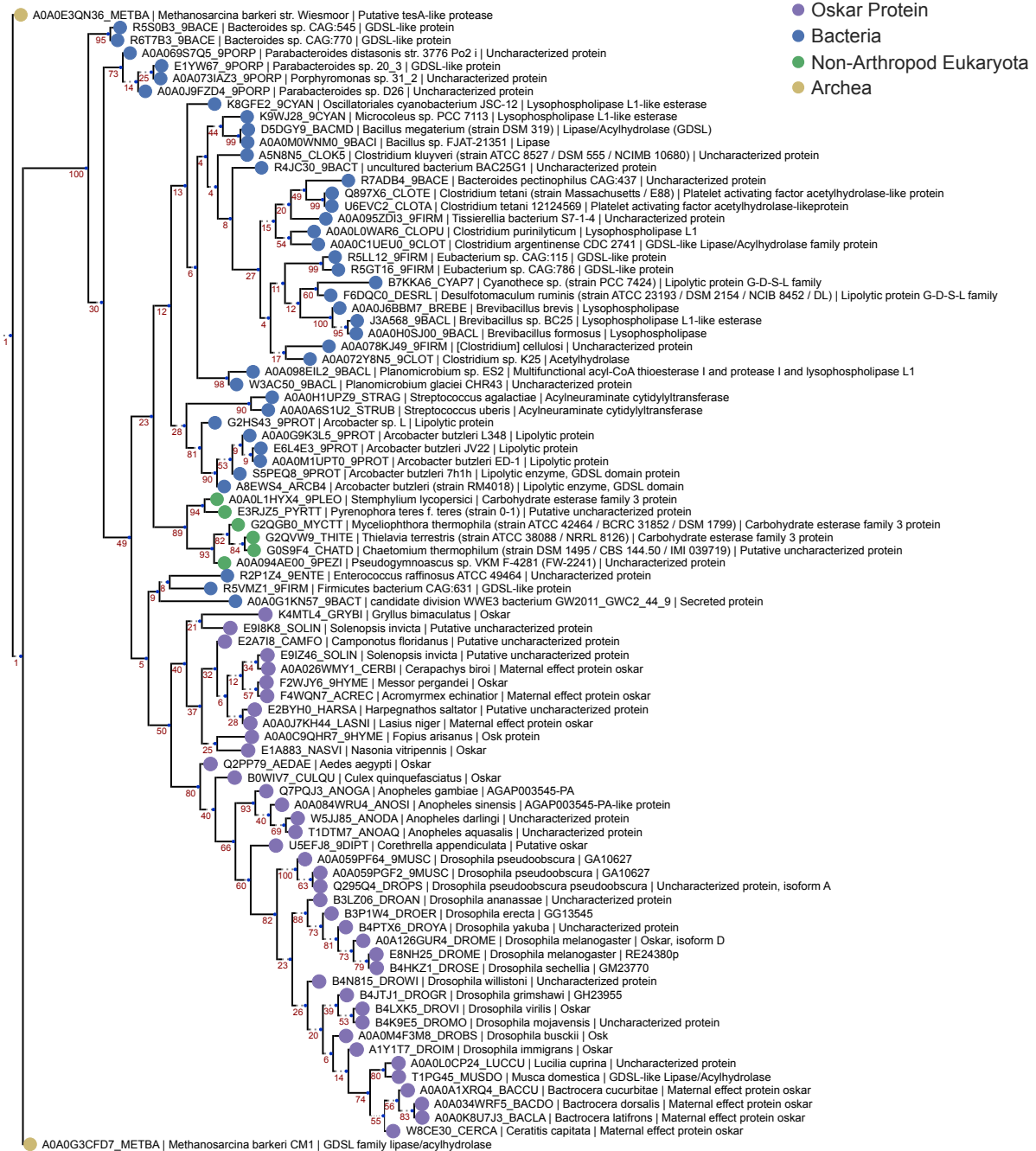
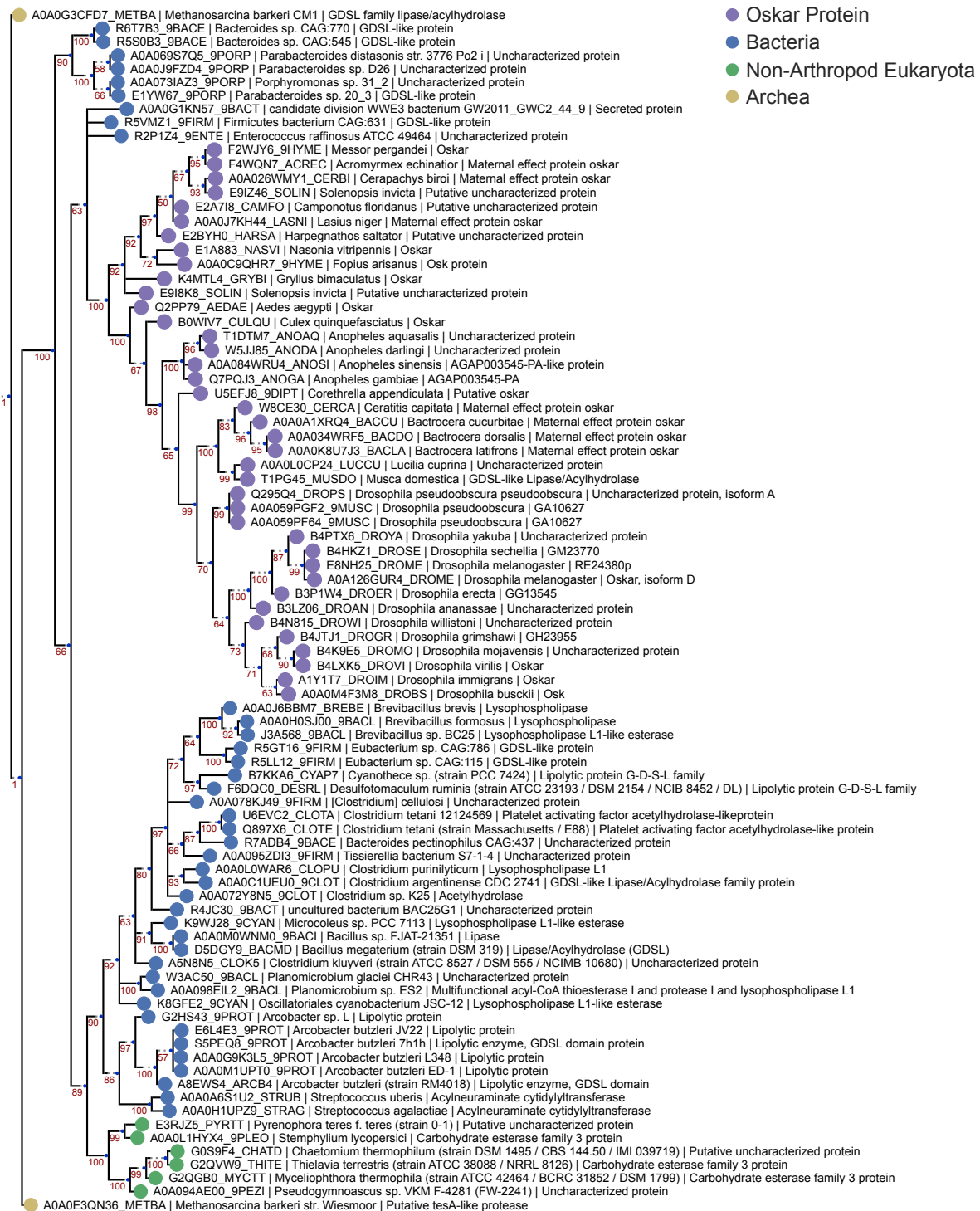


Figure A.4 (following page): OSK Domain Bayesian MUSCLE Tree. Phylogenetic tree of the HMMER sequences hit on the UniProt database using the OSK alignment HMM model. 87 sequences were chosen for analysis as described for Figure A.3. The tree was created using Mr Bayes V3.2.6 using a Mixed model (prset aamodel=Mixed) and a gamma distribution (lset rates=Gamma). The algorithm was allowed to run for 4 million generations to achieve a std < 0.01. See Section 1.4.5 **Phylogenetic Analysis Based on MUSCLE Alignment** for further detail. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure A.4: (continued)



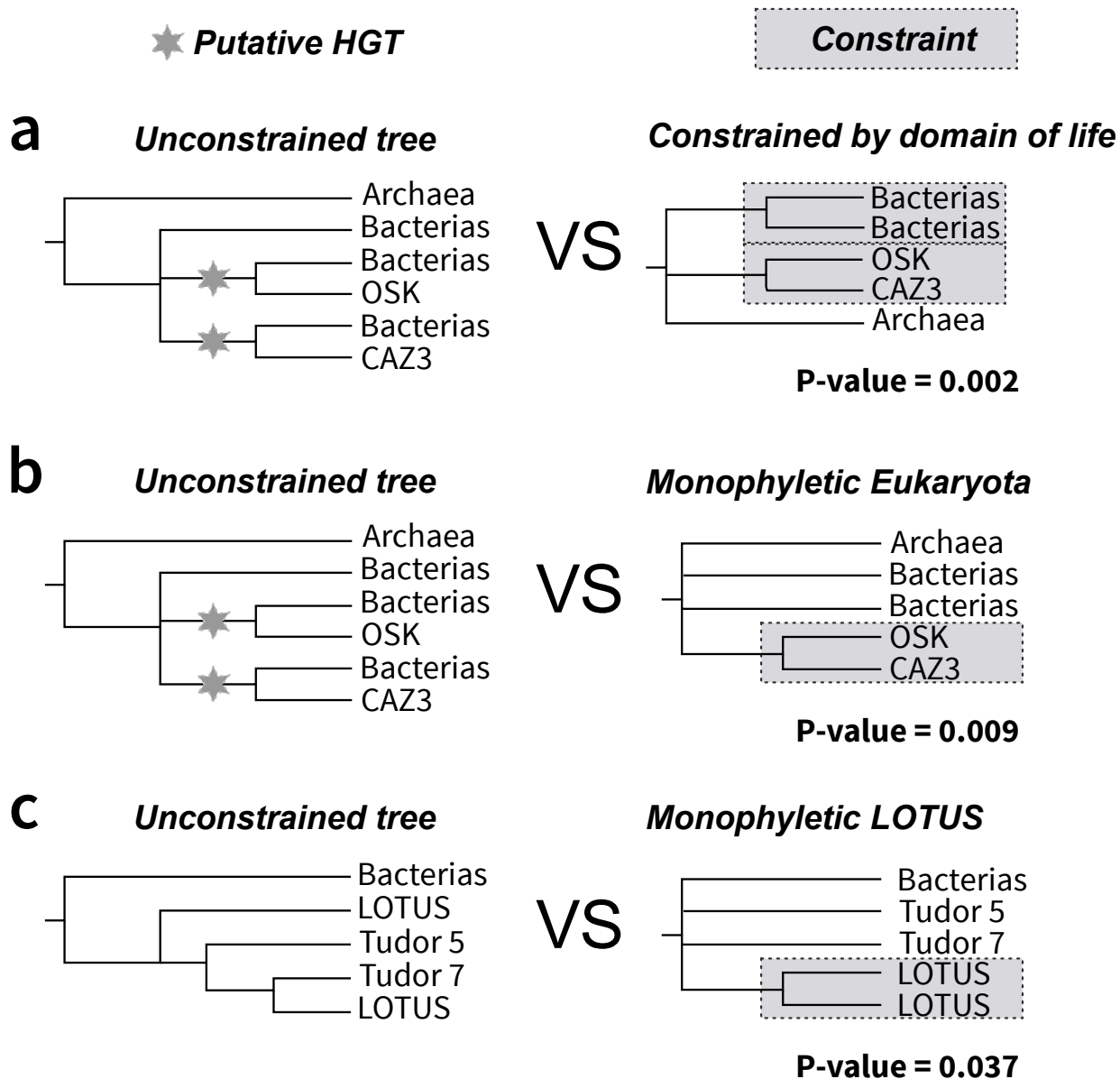


Figure A.5: SOWHAT constrained trees and results. Two trees constrained by alternative relationships that would be expected under vertical transmission of sequences were designed and tested against our result supporting a putative HGT event of the OSK domain. (a) The first tree (right) is constrained by domain of life, requiring bacterial and eukaryotic sequences to be monophyletic, and disallowing sister group relationships of subsets of eukaryotic sequences and bacterial sequences. Our unconstrained tree topology (left) outperformed this topology with a p-value of 0.002 (95% confidence interval upper: 0.007 lower: 0.0002). (b) The second tree requires monophyly of Eukaryota. Our unconstrained tree topology (left) outperformed this topology with a p-value of 0.009 (95% confidence interval upper: 0.017 lower: 0.004). (c) The third tree tested whether the LOTUS domain split observed in the tree generated with the MUSCLE alignment was significantly different from a tree where the LOTUS sequences formed a monophyly. The unconstrained tree (left) outperformed this topology with a p-value of 0.037 (95% confidence interval upper: 0.05 lower: 0.026).

Figure A.6 (following page): LOTUS Domain RaxML PRANK Tree. Phylogenetic tree of the same sequences used for the previous LOTUS trees. The sequences were aligned using PRANK and the tree generated with RaxML as described in Section 1.4.6 **Phylogenetic Analysis Based on PRANK alignment.** Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure A.6: (continued)

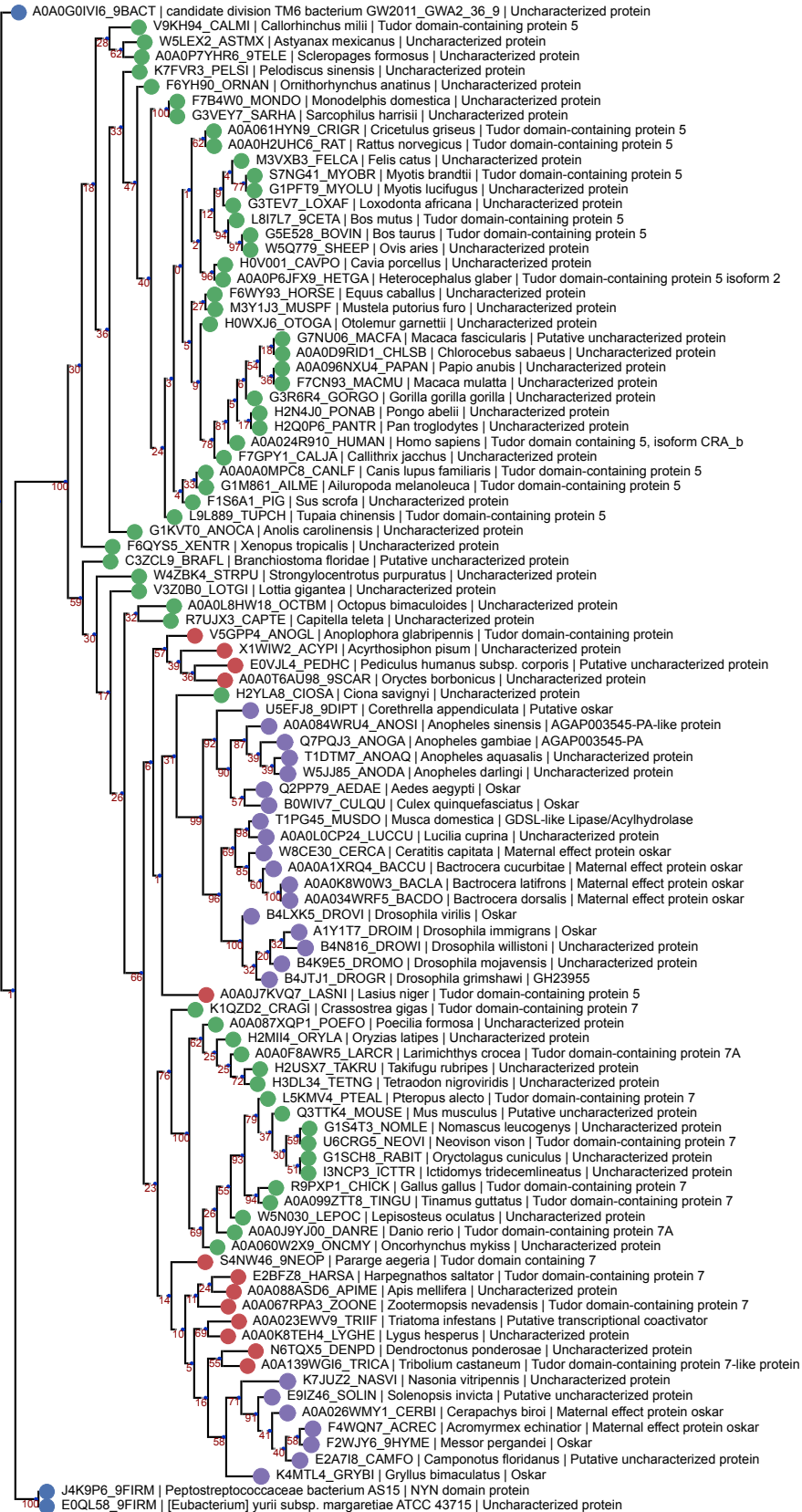
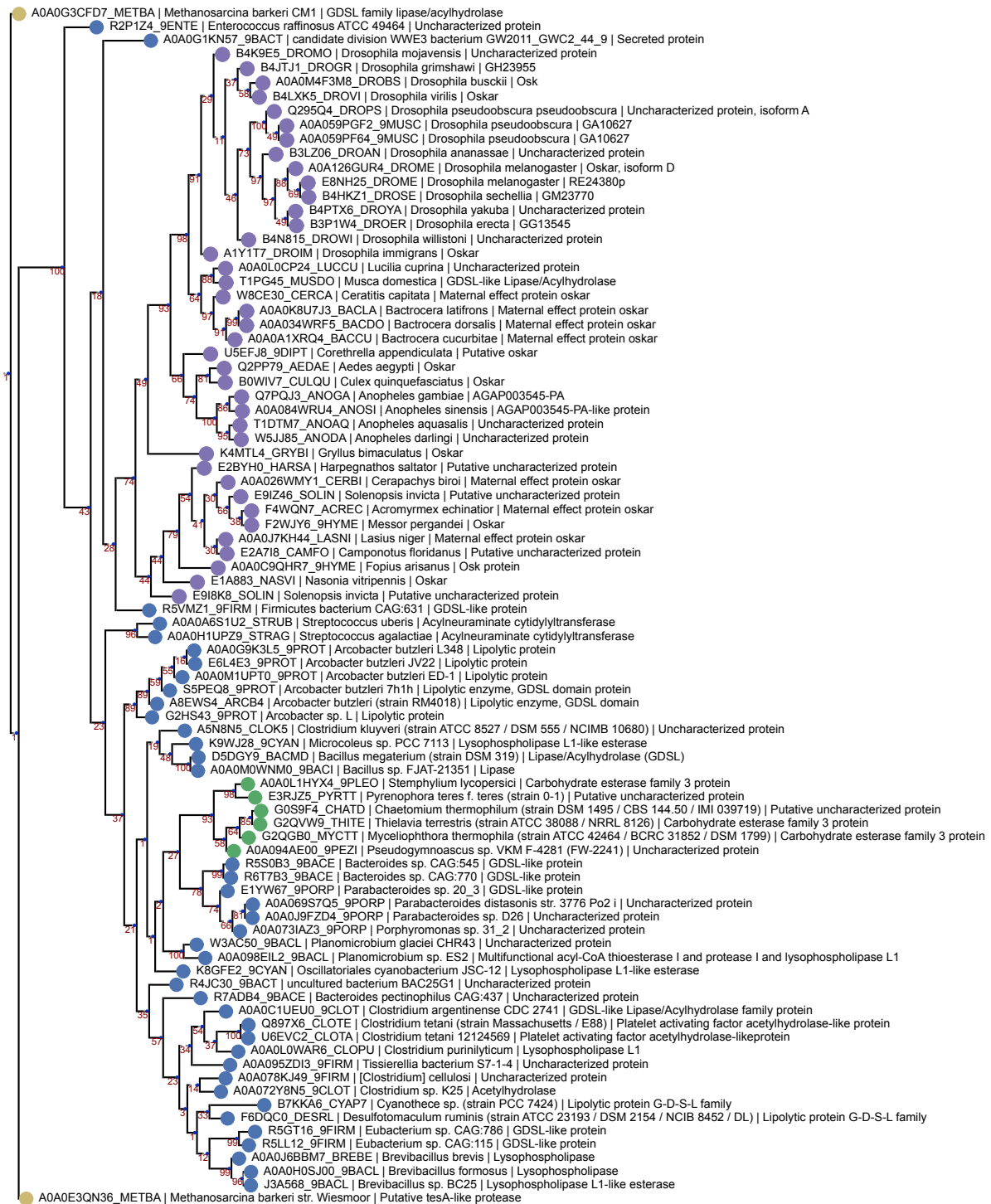


Figure A.7 (following page): OSK Domain RaxML PRANK Tree. Phylogenetic tree of the same sequences used for the previous OSK trees. The sequences were aligned using PRANK and the tree generated with RaxML as described in Section 1.4.6 **Phylogenetic Analysis Based on PRANK alignment**. Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure A.7: (continued)



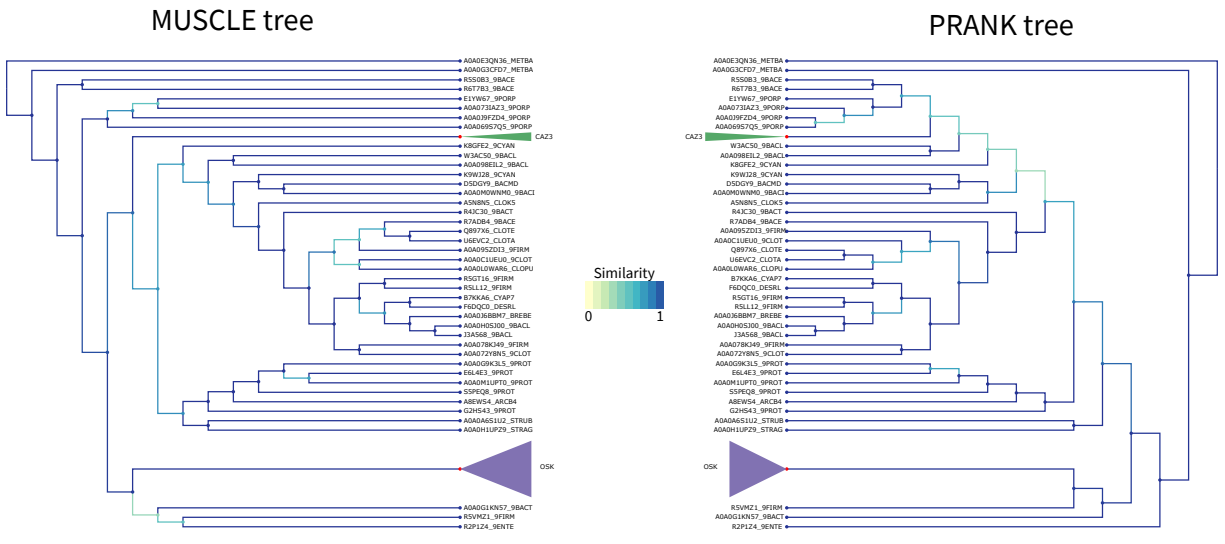


Figure A.8: OSK Tree PRANK Comparison. Comparison of the tree obtained with RaxML starting from the MUSCLE alignment (left) versus the PRANK alignment (right) for the OSK domain. Similarity scores for the branching events are color coded from yellow to blue (see figure color bar legend). The OSK (purple) clade and CAZ3 (green) clade have been colored and compacted for readability as they do not have any internal branching changes. Node color is blue if the leaf is a sequence, and red if this is a compacted group of sequences.

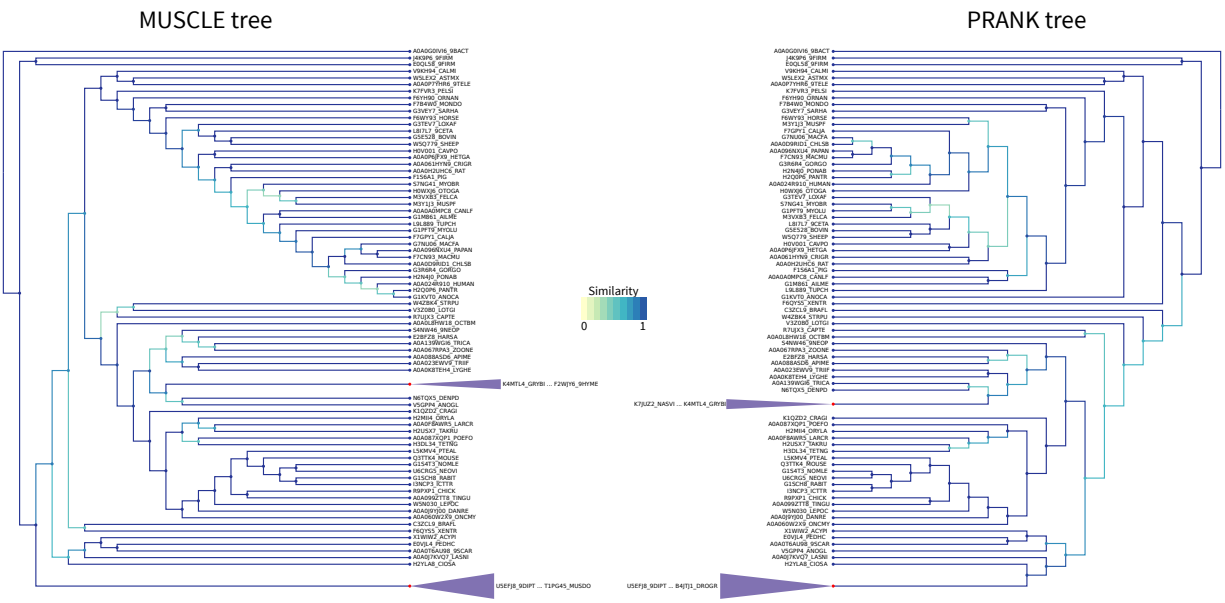


Figure A.9: LOTUS Tree PRANK Comparison. Comparison of the tree obtained with RaxML starting from the MUSCLE alignment (left) versus the PRANK alignment (right) for the LOTUS domain. Similarity scores for the branching events are color coded from yellow to blue (see figure color bar legend). The LOTUS (purple) clades have been colored and compacted for readability as they do not have any internal branching changes. Node color is blue if the leaf is a sequence, and red if this is a compacted group of sequences.

Figure A.10 (following page): LOTUS Domain RaxML T-Coffee Tree. Phylogenetic tree of the same sequences used for the previous LOTUS trees. The sequences were aligned using T-Coffee and the tree generated with RaxML as described in Section 1.4.7 **Phylogenetic Analysis Based on T-Coffee alignment.** Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure A.10: (continued)

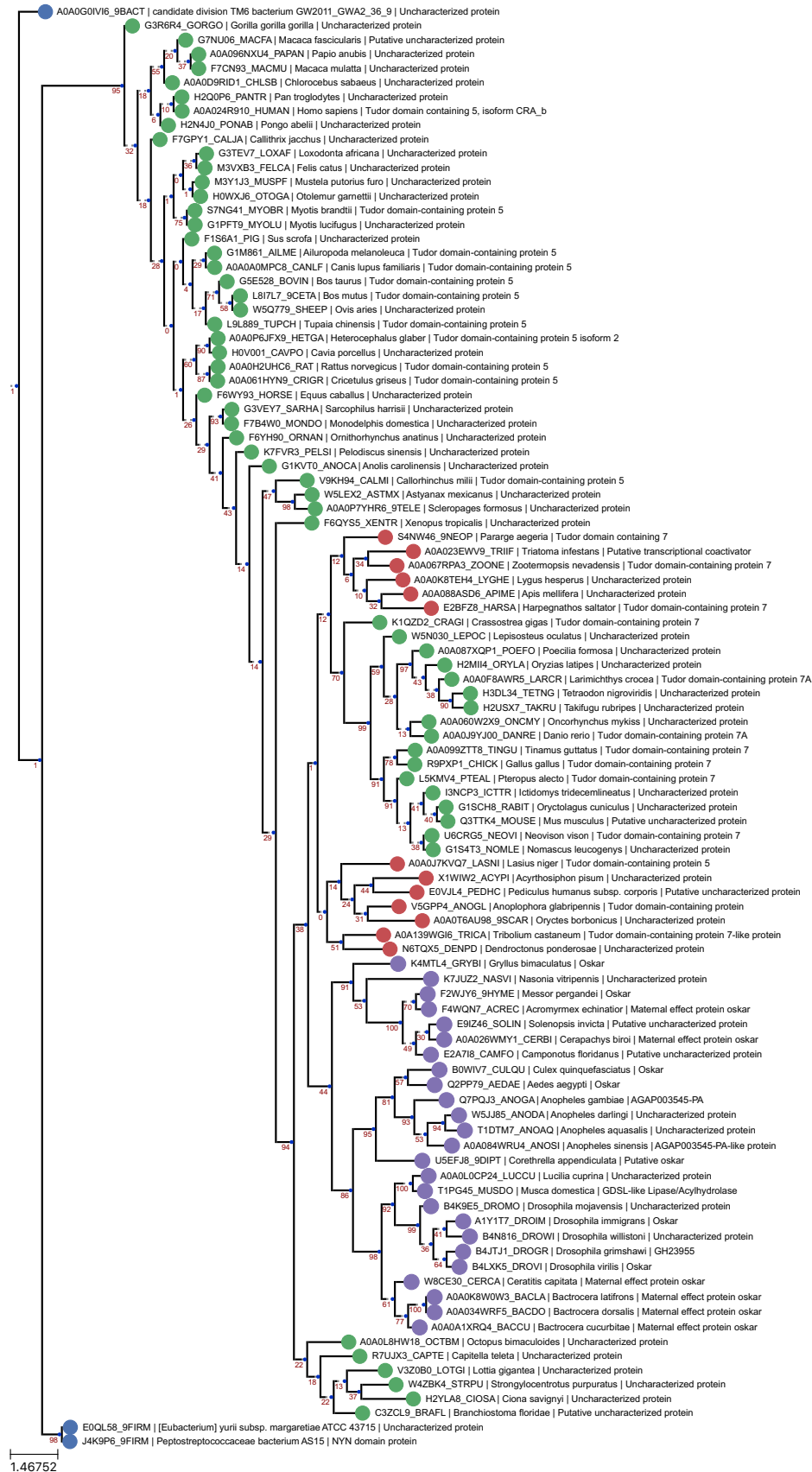
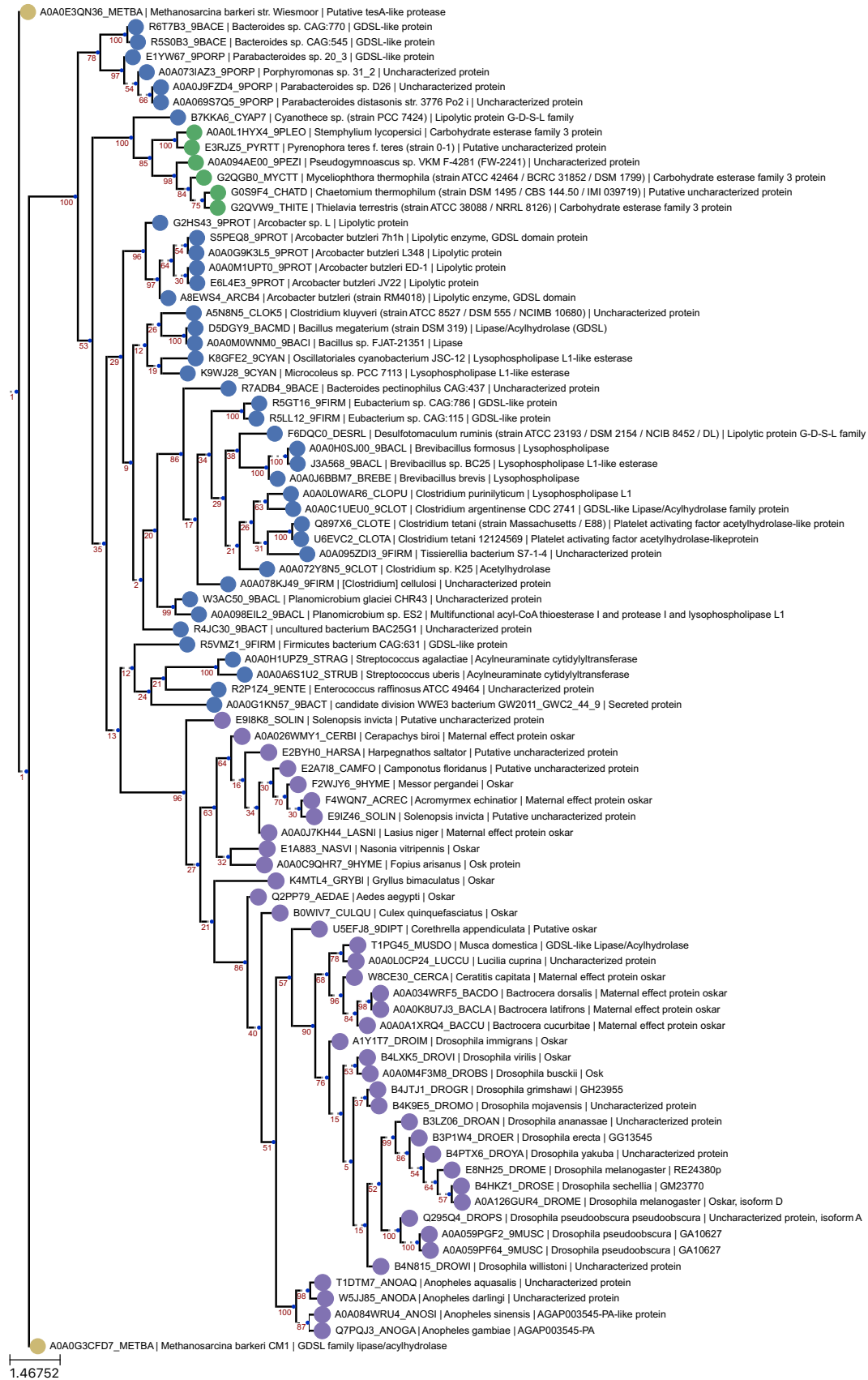


Figure A.11 (following page): OSK Domain RaxML T-Coffee Tree. Phylogenetic tree of the same sequences used for the previous OSK trees. The sequences were aligned using T-Coffee and the tree generated with RaxML as described in Section 1.4.7 **Phylogenetic Analysis Based on T-Coffee alignment.** Sequences are color-coded as follows: Purple = Oskar; Red = Non-Oskar Arthropod; Green = Non-Arthropod Eukaryote; Blue = Bacteria. Names following leaves display the UniProt accession number followed by the species name and the UniProt protein name.

Figure A.11: (continued)



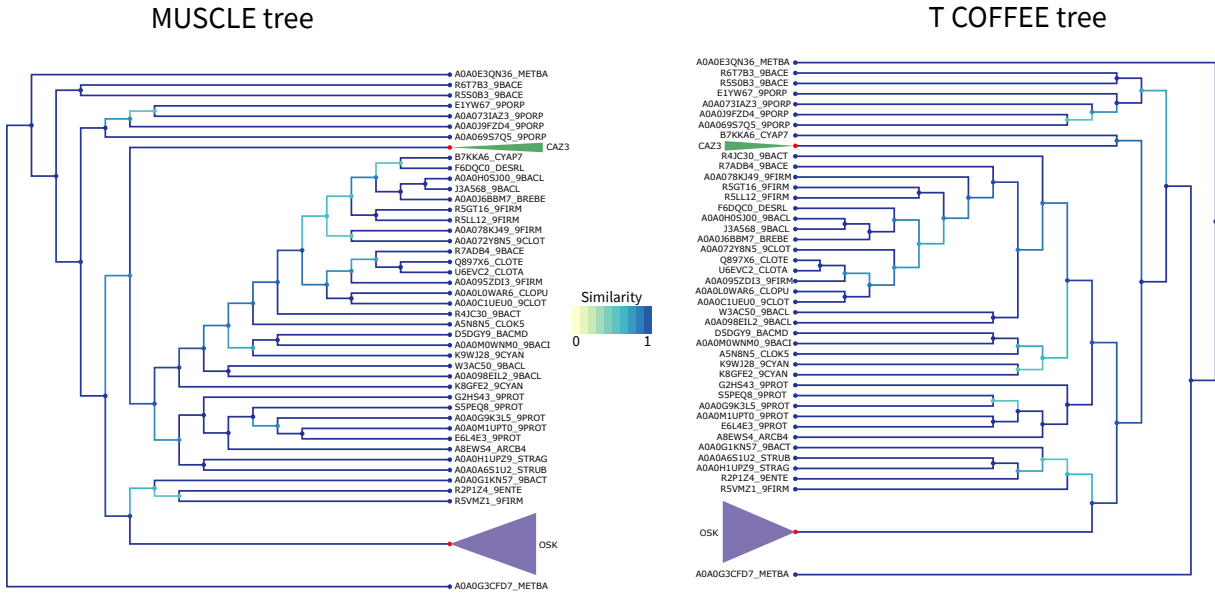


Figure A.12: OSK Tree T-Coffee Comparison. Comparison of the tree obtained with RaxML starting from the MUSCLE alignment (left) versus the T-Coffee alignment (right) for the OSK domain. Similarity scores for the branching events are color coded from yellow to blue (see figure color bar legend). The OSK (purple) clade and CAZ3 (green) clade have been colored and compacted for readability as they do not have any internal branching changes. Node color is blue if the leaf is a sequence, and red if this is a compacted group of sequences.

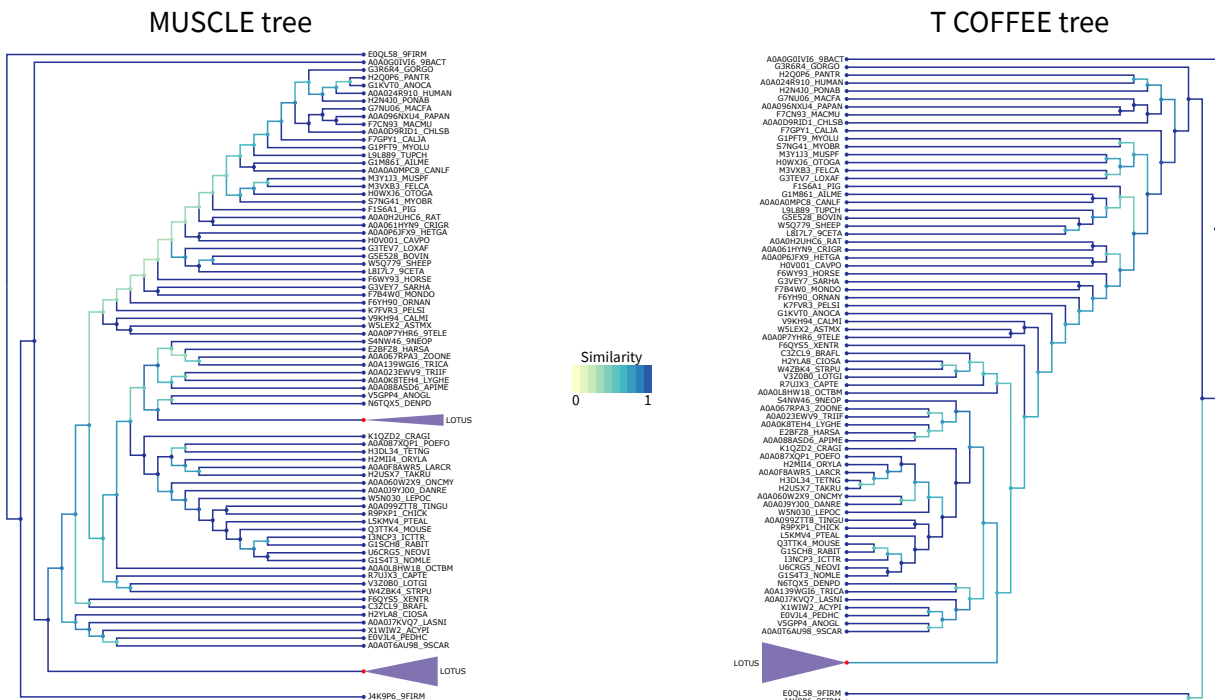


Figure A.13: LOTUS Tree T-Coffee Comparison. Comparison of the tree obtained with RaxML starting from the MUSCLE alignment (left) versus the T-Coffee alignment (right) for the LOTUS domain. Similarity scores for the branching events are color coded from yellow to blue (see figure color bar legend). The LOTUS (purple) clades have been colored and compacted for readability as they do not have any internal branching changes. Node color is blue if the leaf is a sequence, and red if this is a compacted group of sequences.

Table A.1: List of genomes and transcriptomes used for automated oskar search. The table can be accessed here: https://github.com/extavourlab/Oskar_HGT/blob/master/Data/Tables/Supp_Table1.csv

List of genomes and transcriptomes that were downloaded, annotated, and searched for oskar sequences (see “Hidden Markov Model (HMM) generation and alignments of the OSK and LOTUS domains” in Methods). The table reports the database provenance (NCBI genome or TSA, or 1KITE database) and the accession number. The TSA accession ID can be searched using the NCBI TSA browser here: <https://www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSA>.

Table A.2: List of oskar sequences used in the final alignment. The table can be accessed here: https://github.com/extavourlab/Oskar_HGT/blob/master/Data/Tables/Supp%20Table2.csv

List of accession numbers and database provenance of the sequences used in the final alignments of Oskar analysed herein. The table contains the database provenance (Type), the database accession number (ID), the species, family and order, and extraction notes. In the “Annotation” Column, P = homolog identified by pipeline; DB = homolog identified by database annotation. *Sequence recomposed from two transcripts: GBCX01024638.1 and GBCX01024637.

Table A.3: List of sequences and their BLAST results used for phylogenetic analysis of the LOTUS domain. The table can be accessed here: https://github.com/extavourlab/Oskar_HGT/blob/master/Data/Tables/Supp%20Table3.csv

The sequences were obtained by searching the TrEMBL database using hmmsearch and the final HMM generated for LOTUS (Supplementary files: HMM >LOTUS.hmm). Reported are the UniProtID (Accession Number), the Domain and Phylum origin of the sequence, the E-value, score and bias given by hmmsearch, and the description of the target from UniProt. To obtain sequences for each entry, either search UniProt directly (<https://www.uniprot.org/>) or consult the final alignment in Supplementary Files: Alignments >LOTUS_TREE.fasta. Phylum abbreviations: A = Arthropoda; An = Annelida; E = Echinodermata; F = Firmicutes; M = Mollusca; T = Tunicata; V = Vertebrata; ? = unclassified

Table A.4: List of sequences and their BLAST results used for phylogenetic analysis of the OSK domain. The table can be accessed here: https://github.com/extavourlab/Oskar_HGT/blob/master/Data/Tables/Supp%20Table4.csv

The sequences were obtained by searching the TrEMBL database using hmmsearch and the final HMM generated for OSK (Supplementary files: HMM >OSK.hmm). Reported parameters are as described for Supplementary Table S3. To obtain sequences for each entry, either search UniProt directly (<https://www.uniprot.org/>) or consult the final alignment in Supplementary Files: Alignments >OSK_TREE.fasta. Phylum Abbreviations: A = Arthropoda; Ar = Archaea; As = Ascomycota; B = Bacteroidetes; C = Cyanobacteria; Eu = Euryarchaeota; F = Firmicutes; Fu = Fungi; P = Proteobacteria

B

Chapter 2: Supplementary data

This appendix contains the supplementary figures for Chapter 2.

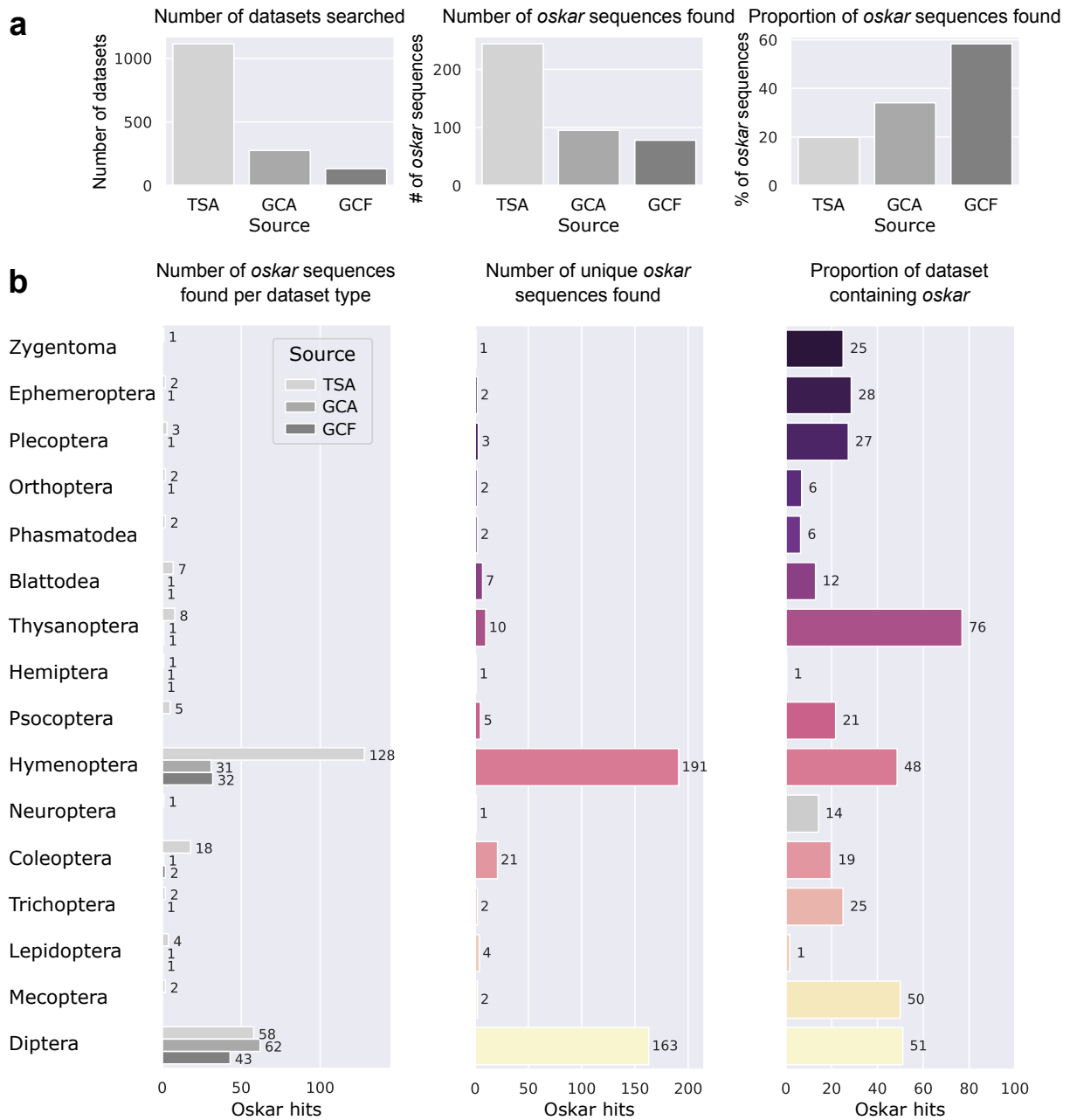


Figure B.1: Summary statistics. In **a**), Overall statistics for each of the three sources of datasets searched, from left to right: The number of datasets searched coming from any of the sources, the number of *oskar* sequences found in each of those datasets, and the proportion of *oskar* sequences found in any of any of the three sources. In **b**), Summary statistics broken down by insect orders. Only orders where an *oskar* sequence was found are shown. From left to right: The number of *oskar* sequences found in each of the three data sources, the total number of *oskar* sequences found, the proportion of *oskar* sequences found.

Figure B.2 (following page): Genome and Transcriptome quality correlation to *oskar* discovery. Shown are box plots of the distribution of multiple genome and transcriptome quality metrics. The distributions were split between the presence and absence of *oskar* in a dataset (found / not found). For each metric, the mean of both distributions was tested for significance using a Mann Whitney U test, and a bar with an * is displayed if the p-value was less than 0.05.

Figure B.2: (continued)

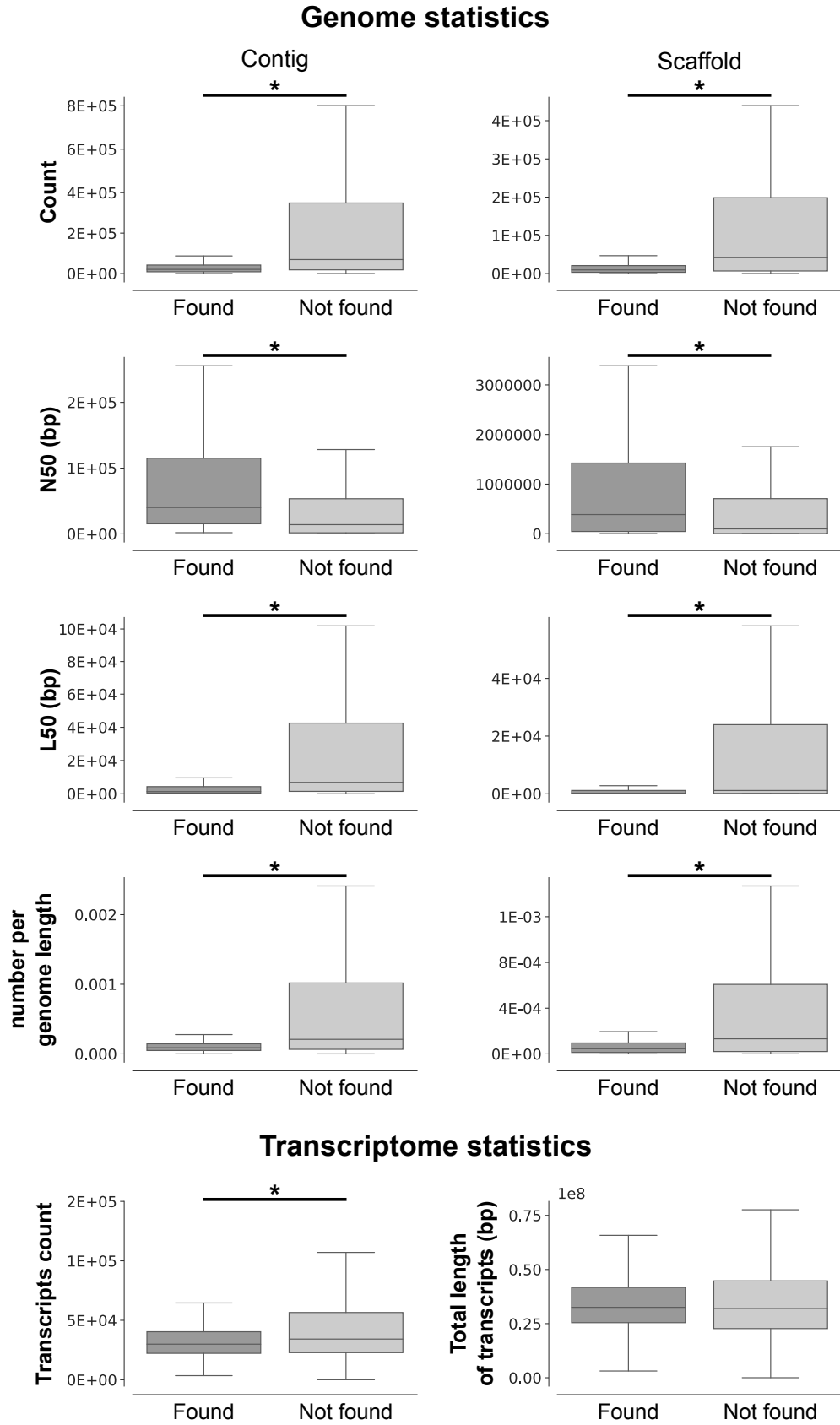
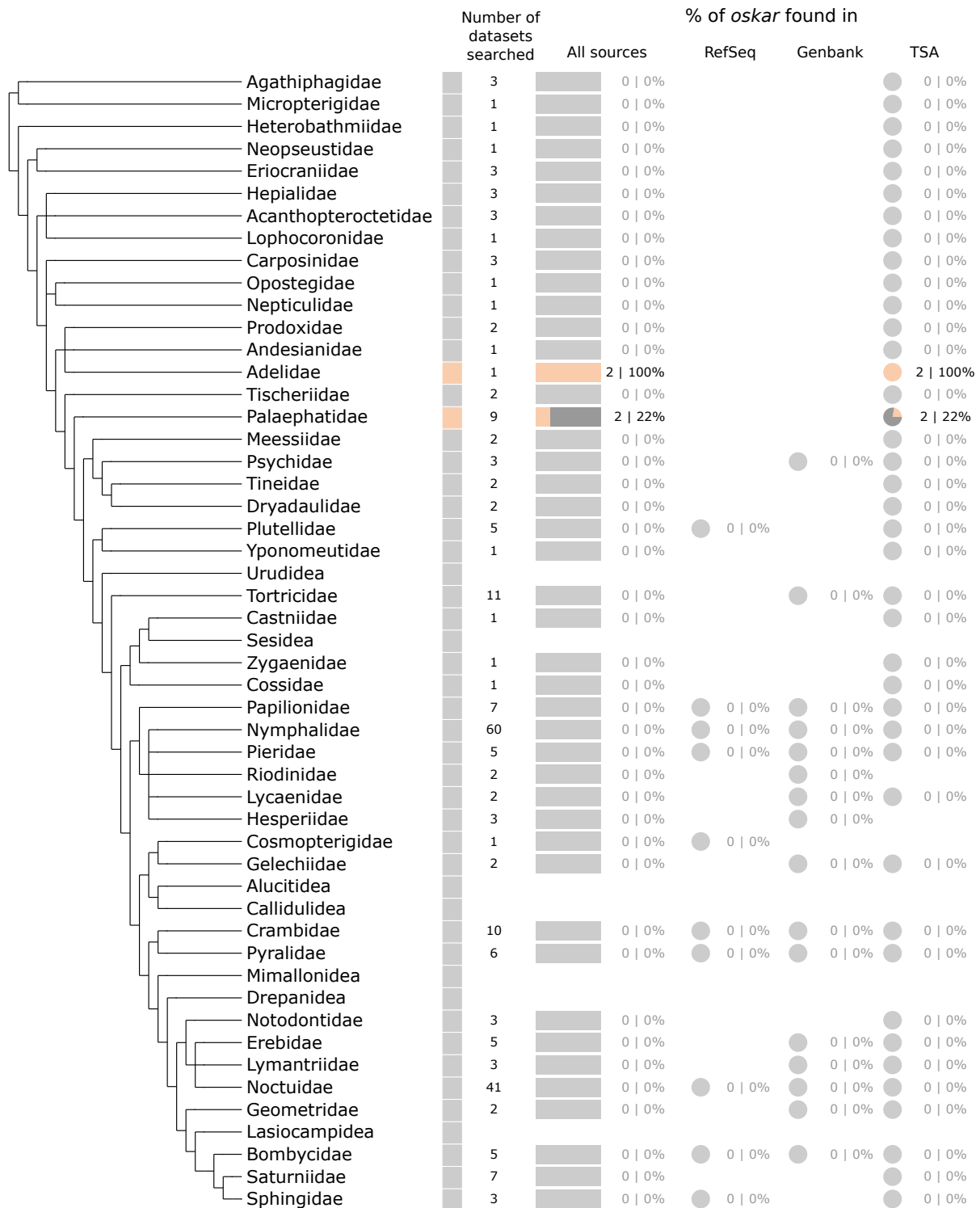


Figure B.3 (following page): Loss of *oskar* in Lepidoptera. Phylogeny of the Lepidopteran extracted from Kawahara et al. ¹. Next to each lepidopteran family are shown summary data regarding the status of *oskar* discovery. In gray, no *oskar* sequences were found, color means some *oskar* sequences were found. From left to right: The number of datasets where we searched for *oskar*, the number of *oskar* sequences found in those datasets, the number of *oskar* sequences found in RefSeq genomes, Genbank genomes and TSA transcriptomes.

Figure B.3: (continued)



Insect Order	Source	Number of dataset	Total hits	Filtered hits	Proportion with oskar found
Archaeognatha	GCA	1	0	0	0%
Archaeognatha	TSA	2	0	0	0%
Blattodea	GCA	3	1	1	33%
Blattodea	GCF	2	0	0	0%
Blattodea	TSA	51	7	7	14%
Coleoptera	GCA	12	1	1	8%
Coleoptera	GCF	9	3	2	22%
Coleoptera	TSA	86	31	14	16%
Collembola	TSA	9	0	0	0%
Dermaptera	TSA	7	0	0	0%
Diptera	GCA	115	63	60	52%
Diptera	GCF	43	58	43	100%
Diptera	TSA	162	72	58	36%
Embioptera	TSA	5	0	0	0%
Ephemeroptera	GCA	2	0	0	0%
Ephemeroptera	TSA	5	1	1	20%
Grylloblattodea	TSA	2	0	0	0%
Hemiptera	GCA	18	0	0	0%
Hemiptera	GCF	12	0	0	0%
Hemiptera	TSA	192	1	0	0%
Hymenoptera	GCA	52	32	30	58%
Hymenoptera	GCF	47	36	32	68%
Hymenoptera	TSA	301	157	128	43%
Lepidoptera	GCA	80	0	0	0%
Lepidoptera	GCF	17	0	0	0%
Lepidoptera	TSA	135	24	4	3%
Mantodea	TSA	13	0	0	0%

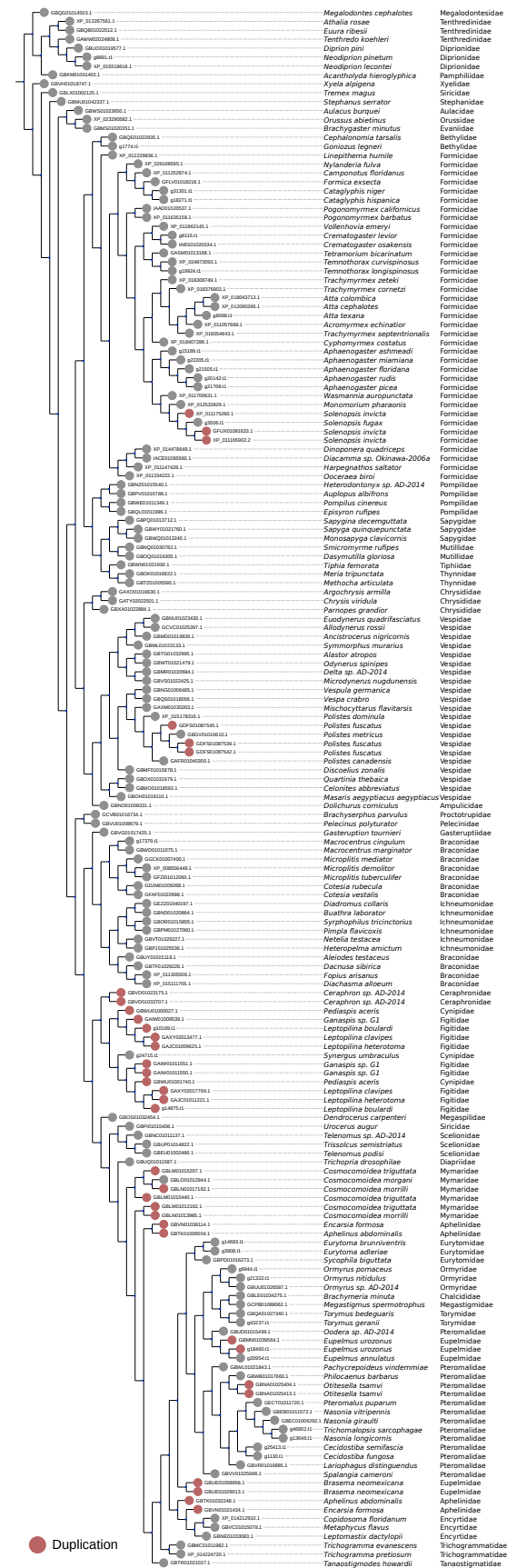
Insect Order	Source	Number of dataset	Total hits	Filtered hits	Proportion with <i>oskar</i> found
Mantophasmatodea	TSA	2	0	0	0%
Mecoptera	TSA	4	2	2	50%
Megaloptera	TSA	3	0	0	0%
Neuroptera	TSA	7	1	1	14%
Odonata	GCA	2	0	0	0%
Odonata	TSA	7	0	0	0%
Orthoptera	GCA	3	0	0	0%
Orthoptera	TSA	28	2	2	7%
Phasmatodea	GCA	13	0	0	0%
Phasmatodea	TSA	31	6	2	6%
Phthiraptera	GCF	1	0	0	0%
Phthiraptera	TSA	7	0	0	0%
Plecoptera	GCA	3	0	0	0%
Plecoptera	TSA	8	3	3	38%
Psocoptera	TSA	23	5	5	22%
Raphidioptera	TSA	3	0	0	0%
Siphonaptera	GCF	1	0	0	0%
Siphonaptera	TSA	4	0	0	0%
Strepsiptera	GCA	1	0	0	0%
Strepsiptera	TSA	2	0	0	0%
Thysanoptera	GCA	1	1	1	100%
Thysanoptera	GCF	1	1	1	100%
Thysanoptera	TSA	11	10	8	73%
Trichoptera	GCA	3	1	1	33%
Trichoptera	TSA	7	2	2	29%
Zoraptera	TSA	2	0	0	0%
Zygentoma	TSA	4	1	1	25%

Insect Order	Source	Number of dataset	Total hits	Filtered hits	Proportion with <i>oskar</i> found
Crustacea	TSA	168	0	0	0%
Crustacea	GCF	1	0	0	0%
Crustacea	GCA	11	0	0	0%

Table B.1: Number of *oskar* sequence found per order and per data source. Each line corresponds to an order and a data source: GCF, GCA, TSA. The number of total hits is reported as well as the number of hits after the filtration algorithm described in the Methods is applied. Finally, the proportion of *oskar* sequences found, defined as the number of datasets with a positive hit divided by the total number of datasets searched, is reported.

Figure B.4 (following page): Complete hymenopteran Oskar phylogeny. Phylogenetic tree of all hymenopteran Oskar sequences inferred using RaxML with 100 bootstrap. Branch length normalized to only show the topology. Each leaf is an Oskar ortholog. In gray, only one Oskar sequence was found in this species, in red duplicated Oskars sequences (sequence similarity < 80%).

Figure B.4: (continued)



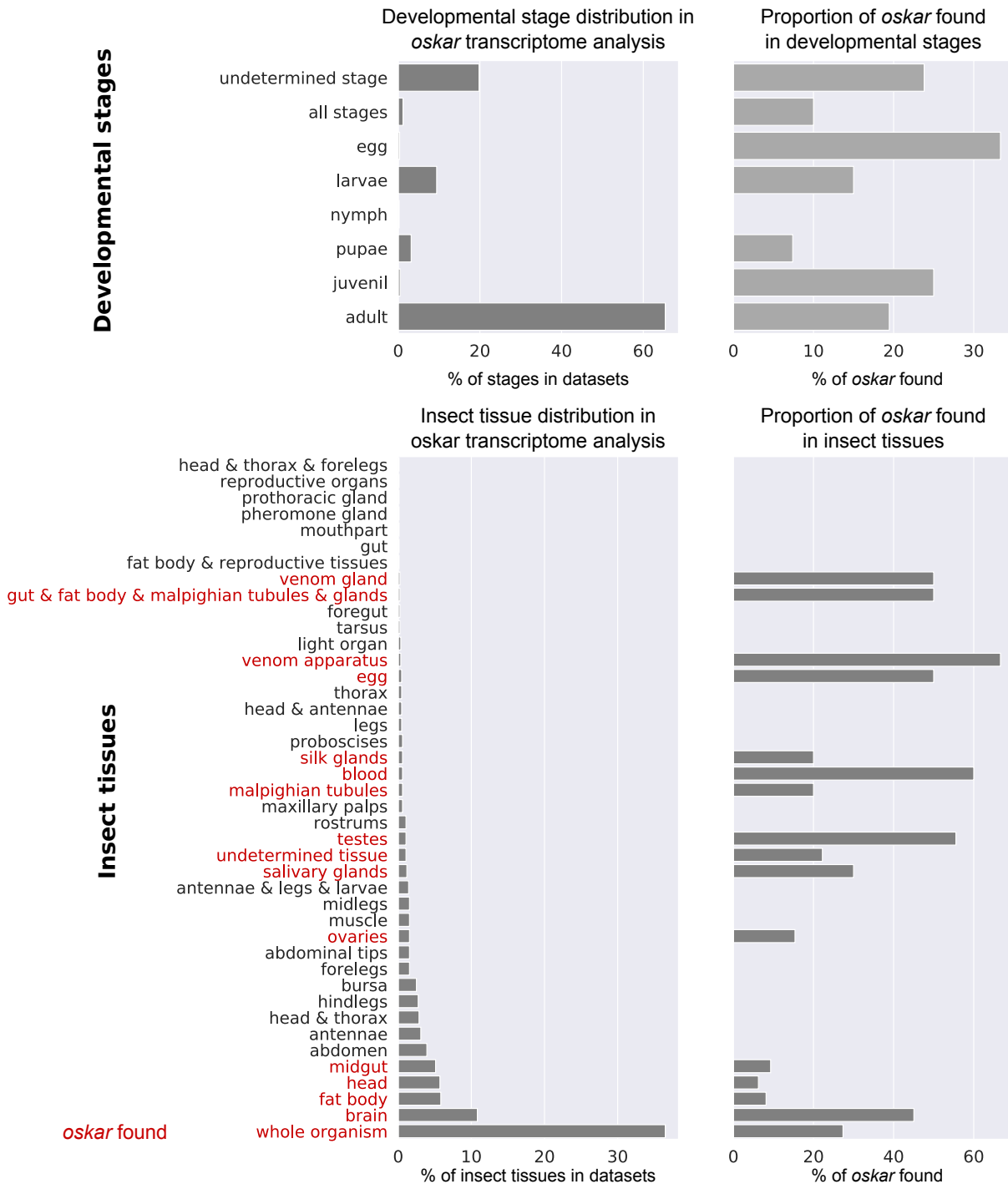


Figure B.5: Tissue and Stage metadata analysis of *oskar* presence in transcriptomes datasets. On the left, the proportion of datasets analyzed with the corresponding stage or tissue type. On the right, the proportion of datasets with a given stage or tissue type where *oskar* was detected. In red, tissue types where *oskar* was detected.

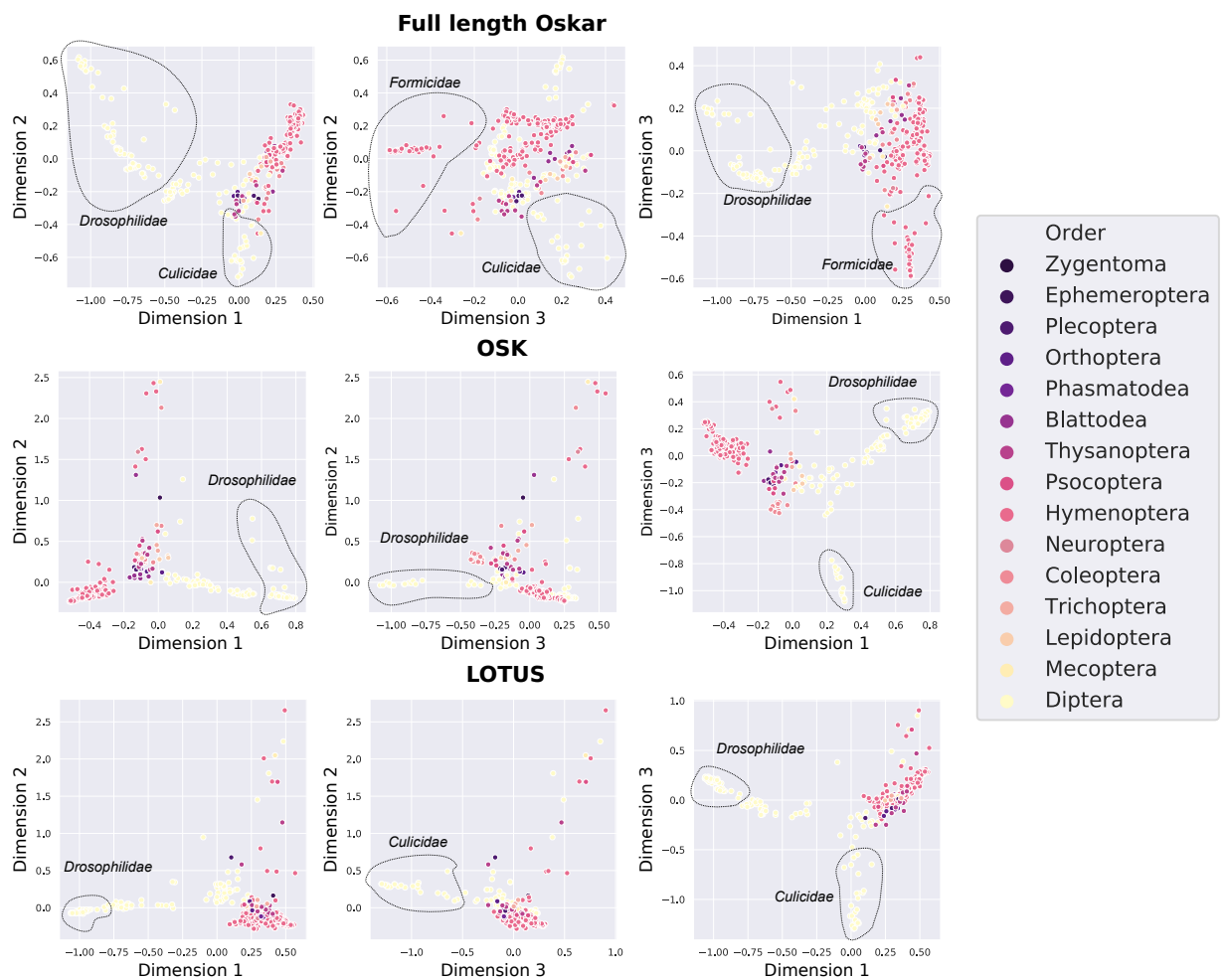


Figure B.6: Multiple Correspondence Analysis (MCA) of Oskar, OSK and LOTUS. MCA analysis of trimmed (30% occupancy) alignments for Oskar, OSK and LOTUS colored by order (see legend). The alignment was projected onto the first three main MCA dimensions. Each dot corresponds to one sequence.

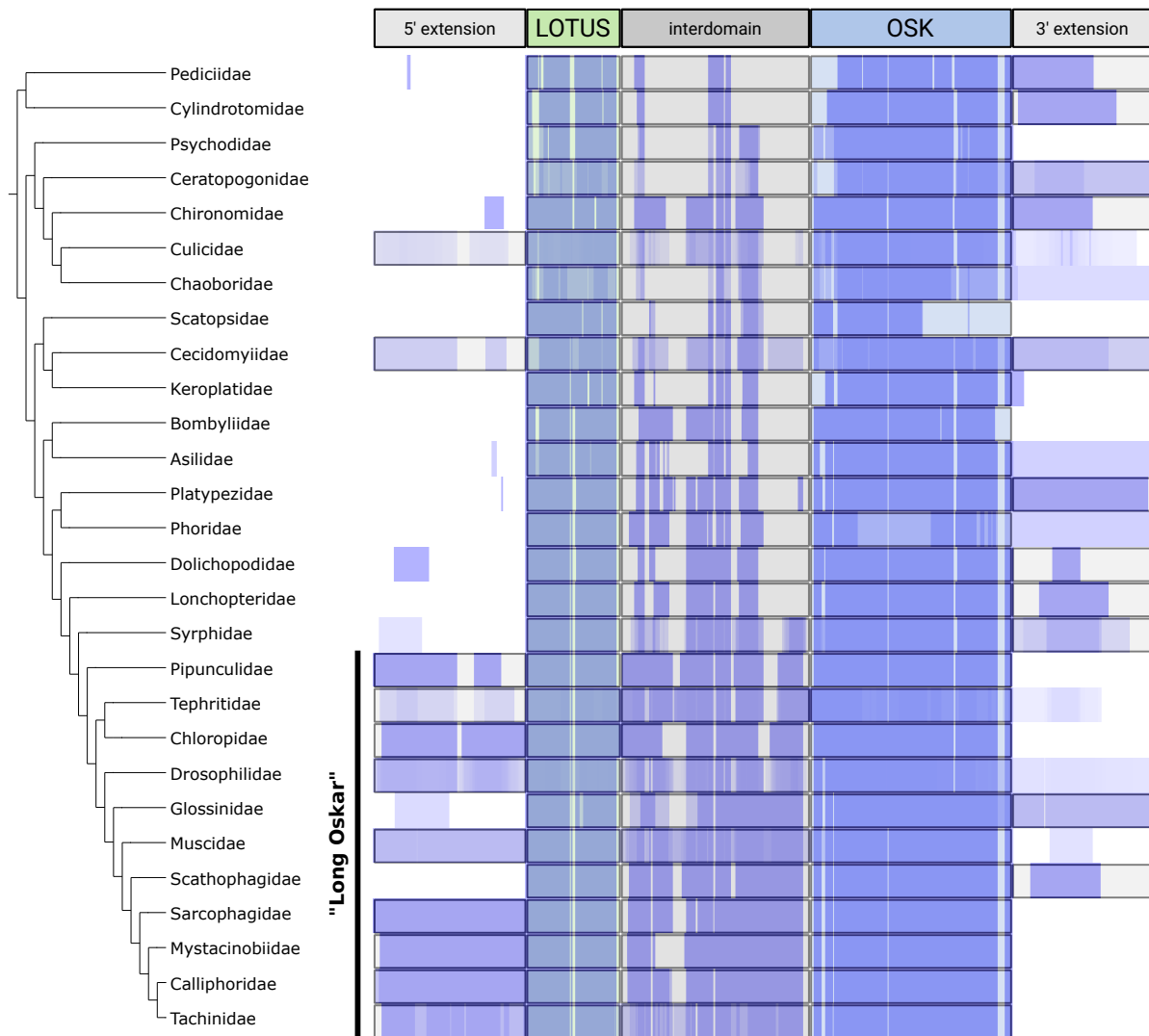


Figure B.7: Evolution of the structure of Oskar in Diptera. On the left is the dipteran phylogeny from Wiegmann et al. ², Maddison et al. ³. At the top is a schematic representation of the Oskar structure. In blue is a heatmap showing the overall occupancy of a position in the Oskar alignment trimmed for at least 10% overall occupancy at a position. For each Dipteran family, the occupancy at a position is defined as: number of non gap Amino Acids / Number of sequences in that family. If a 3' or 5' extension was detected in a family, a transparent box is shown overlaid. Finally, the Long Oskar isoform is shown on the left of the families where it was detected.

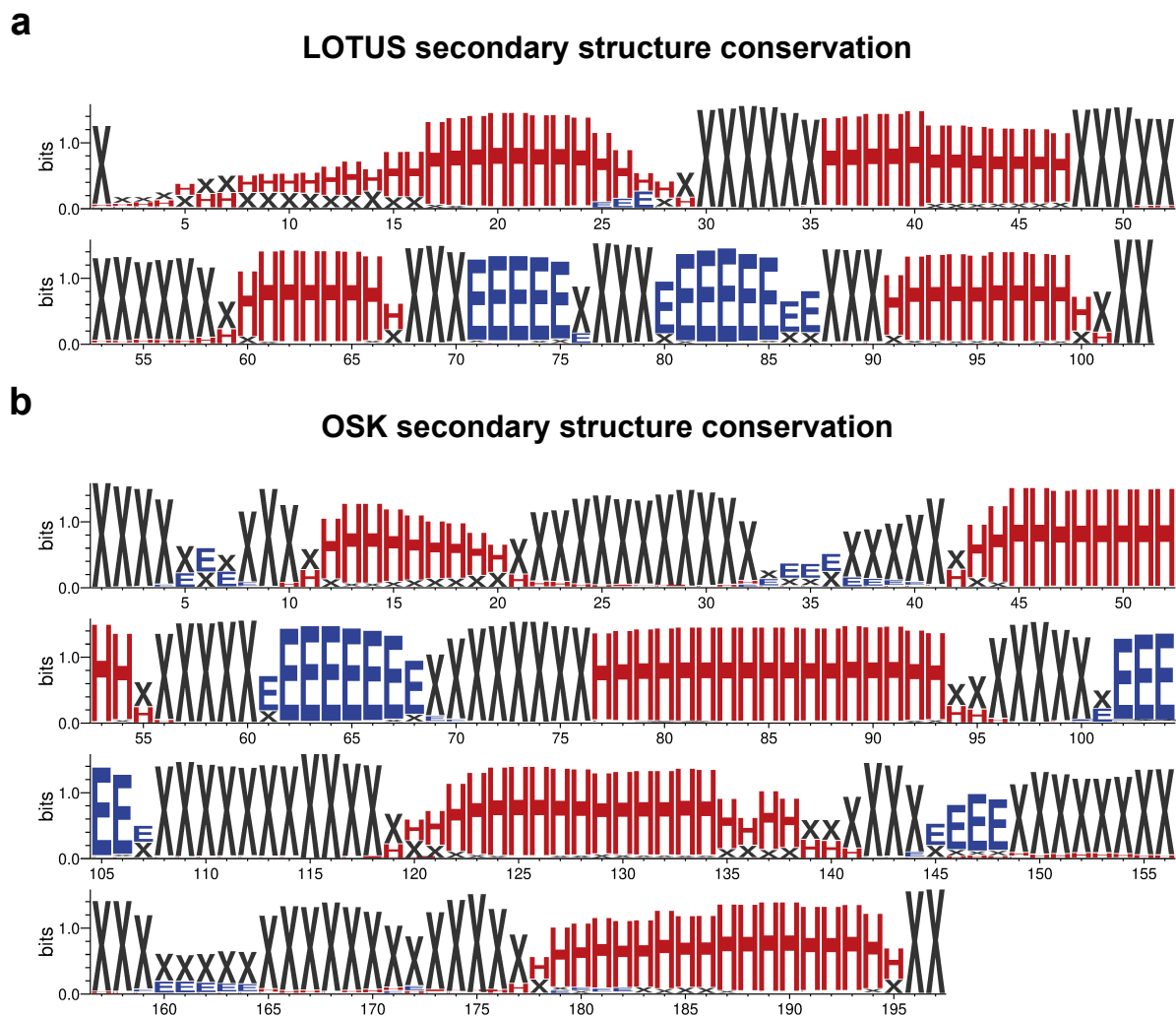


Figure B.8: Oskar domains secondary structure conservation. Sequence logo of Jpred4 predictions for LOTUS and OSK domains showing the conservation of secondary structures. Represented are logos computed with WebLOGO⁴. The height of each letter represents that states (X H or B) conservation throughout the alignment in bits. X (black) are unfolded amino acids, H (red) are α helices and E (blue) are β sheets. In **a**), prediction for the LOTUS domain. In **b**), prediction for the OSK domain.

References

- [1] A. Y. Kawahara, D. Plotkin, M. Espeland, K. Meusemann, E. F. A. Toussaint, A. Donath, F. Gimnich, P. B. Frandsen, A. Zwick, M. Dos Reis, J. R. Barber, R. S. Peters, S. Liu, X. Zhou, C. Mayer, L. Podsiadlowski, C. Storer, J. E. Yack, B. Misof, and J. W. Breinholt. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl. Acad. Sci. U. S. A.*, 116(45):22657–22663, November 2019.
- [2] B. M. Wiegmann, M. D. Trautwein, I. S. Winkler, N. B. Barr, J.-W. Kim, C. Lambkin, M. A. Bertone, B. K. Cassel, K. M. Bayless, A. M. Heimberg, B. M. Wheeler, K. J. Peterson, T. Pape, B. J. Sinclair, J. H. Skevington, V. Blagoderov, J. Caravas, S. N. Kutty, U. Schmidt-Ott, G. E. Kampmeier, F. C. Thompson, D. A. Grimaldi, A. T. Beckenbach, G. W. Courtney, M. Friedrich, R. Meier, and D. K. Yeates. Episodic radiations in the fly tree of life. *Proc. Natl. Acad. Sci. U. S. A.*, 108(14):5690–5695, April 2011.
- [3] D. R. Maddison, K.-S. Schulz, and W. P. Maddison. The tree of life web project. *Zootaxa*, 1668(1):19–40, 2007.
- [4] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14(6):1188–1190, June 2004.



Chapter 3: Supplementary data

This appendix contains the supplementary figures for the Chapter 3.

Table C.1: Tabulation of raw data and analysis for every gene in the screen.

The content of this table can be downloaded on [GitHub](#).

This table contains a summary representation of the data generated by the three screens as well as results from the analysis. Each line corresponds to an independent measurement of a particular RNAi line. Some genes which did not pass the first filter of $|Z_{gene}| > 1$ in the *hpo[RNAi]* Egg Laying screen where then predicted as connectors, therefore they have two entries as they have been independently measured again. The Z scores have been rounded up to 4 significant digits in this table and the Centrality metrics rounded up to 10 significant digits due to their low values, but the full values for both are available in the raw data files provided in the supplementary files in Data/Screens for the Z scores and Results for the centrality values. Moreover, this is a summary table and does not contain values for controls as well as batch numbers, all are available in the supplementary files in Data/Screens.

- **FbID**: FlyBase ID of the tested gene.

- **CG number**: CG Number of the tested gene.

- **NAME**: Common name (as per FlyBase nomenclature) of the gene if existing, else it is a -.

- **SYMBOL**: Symbol (as per FlyBase nomenclature) of the gene if existing, else CG number

- **[ScreenName]_[Variable]_(Metric)_Count**: Within the screen [ScreenName], the count of the measured variable [Variable].

Optional: (*metric*) will indicate if a particular operation was done over the data, such as sum, mean or standard deviation.

e.g. [HippoRNAi_EggL]_[Day_4_Egg]_Count is the count of eggs, on day 4, of the *hpo[RNAi]* Egg Laying screen.

- **[ScreenName]_[Variable]_(Metric)_Zscore**: Within the screen [ScreenName], the Z score of the measured variable [Variable] as calculated to batch control. Optional: (*metric*) will indicate if a particular operation was done over the data, such as sum, mean or standard deviation.

e.g. [EggL]_[All_Days_Egg]_(Sum)_Zscore is the Z score of the sum of eggs count, of the Egg Laying screen.

- **PIN_[Metric]_centrality**: Within the PIN used in this study, the calculated centrality value for the metric [Metric].

- **[SubNetworkName]_Network**: Presence of absence of a gene in the sub-network [SubNetworkName]. If the gene is in the module, this value is True, if it is absent it is False. (An exception is made for the Meta Network displayed in Figure 6 where instead of True/False, the group assignment I-VII is written)

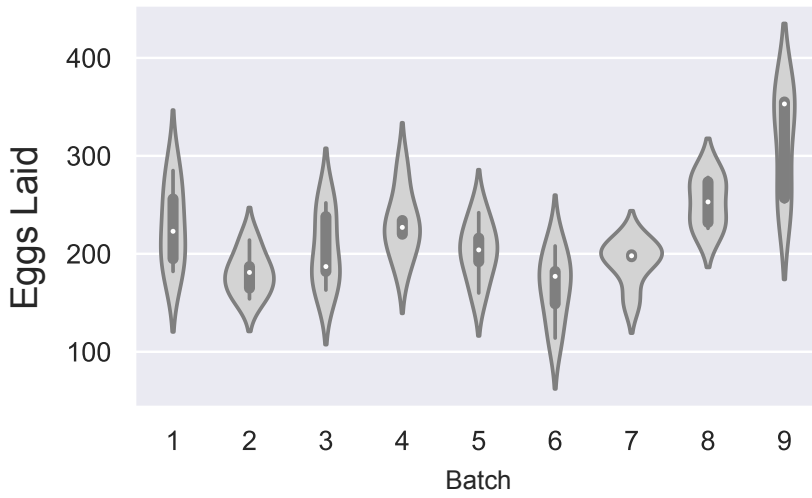
- **[SubNetworkName]_Connector**: Status of a gene in the sub-network [SubNetworkName] as a connector. If True, the gene is a connector, else if False, the gene is not a connector.

- **[PathwayName]_Pathway**: Participation of a gene to the signalling pathway [PathwayName]. If the gene participates in the pathway the value is 1, else it is 0.

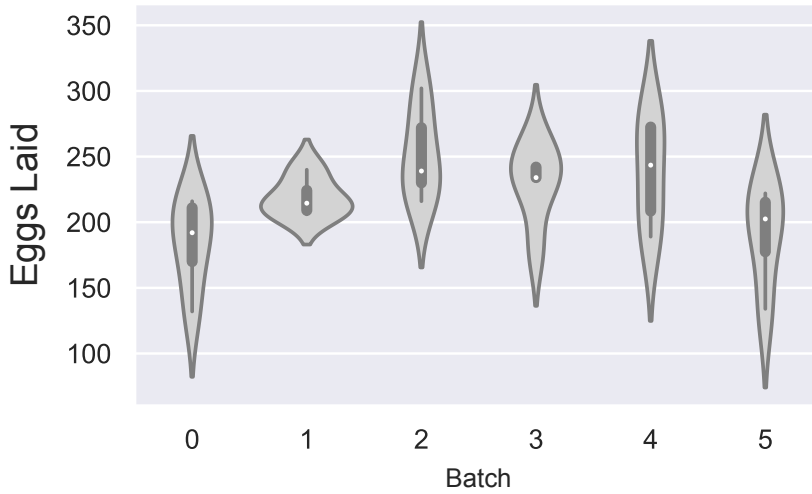
Figure C.1 (following page): Violin plots of egg laying and ovariole number of controls in each screen batch. a) Distribution of number of eggs laid by five replicates of three *tj:Gal4>hpo[RNAi]* females over five days for each batch. **b)** Distribution of number of eggs laid by five replicates of three *tj:Gal4* females over five days for each batch. **c)** Distribution of number of ovarioles per ovary in 20 ovaries from ten *tj:Gal4>hpo[RNAi]* females in each batch.

Figure C.1: (continued)

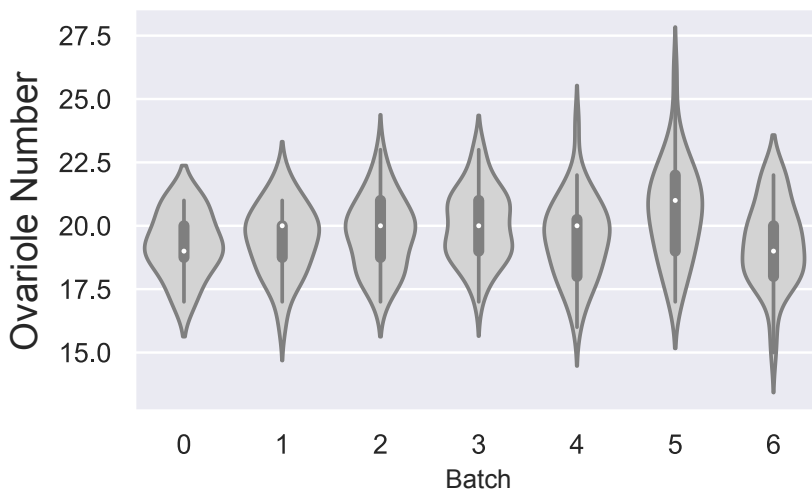
a. *hpo*[RNAi] Egg Laying



b. Egg Laying



c. *hpo*[RNAi] Ovariole Number



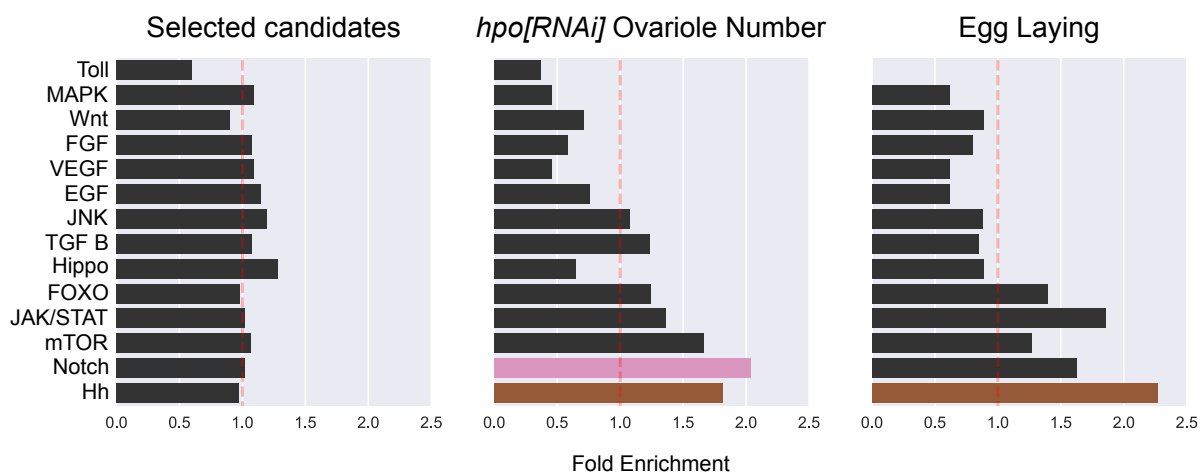


Figure C.2: Enrichment/depletion analysis of the 273 signalling pathway genes above the threshold $|Z_{\text{gene}}| > 1$ (Figure 3.1a) against all signalling candidates. We also measured the enrichment/depletion of positive signalling candidate genes in the *hpo[RNAi]* Ovariole (Figure 1c) and Egg Laying (Figure 1b) screens from the 273 genes tested in those screens.

Centrality	<i>hpo[RNAi] Ovariolo Number</i>	<i>hpo[RNAi] Egg Laying</i>	Egg Laying
Betweenness	0.603	0.57	0.586
EigenVector	0.632	0.573	0.586
Closeness	0.612	0.551	0.588
Degrees	0.615	0.592	0.599

Table C.2: Area under the curve (AUC) of ROC curves. AUC values for the ROC curves for each centrality measure for the three screens (Figure 4a). AUC values range from 0 to 1. A score above 0.5 indicates a positive correlation between the continuous variable (centrality) and the binary variable (above or below the Z score threshold). A score of 0.5 or less indicates no correlation between the variables.

Figure C.3 (following page): Comparison of egg laying candidate genes by pathway. Z_{gene} of egg laying of adult females of $tj>hpo[RNAi], candidate[RNAi]$ plotted against Z_{gene} of egg laying of $tj>candidate[RNAi]$ adult females displayed by pathway. Contour plots indicate a 2D gaussian kernel density estimation.

Figure C.3: (continued)

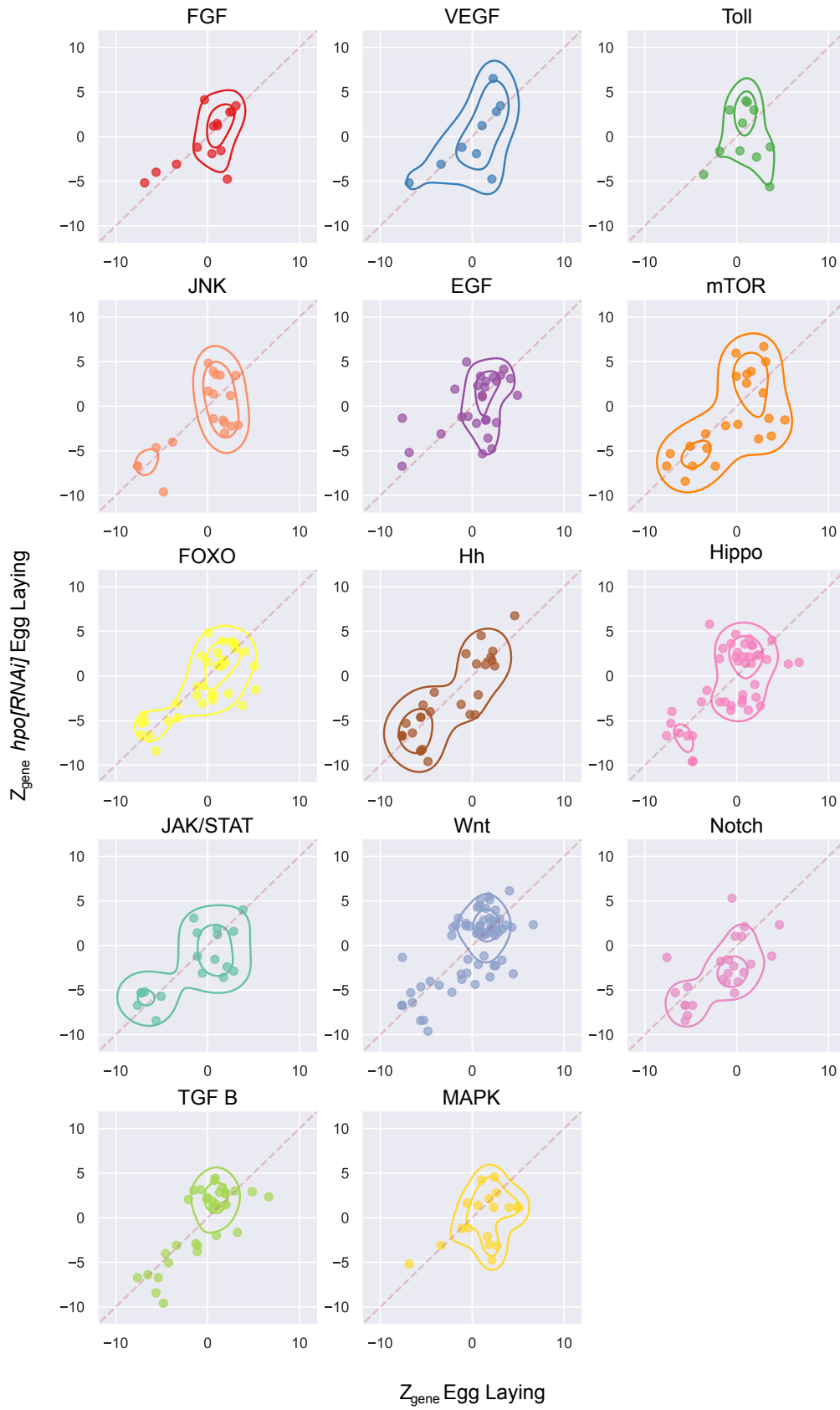
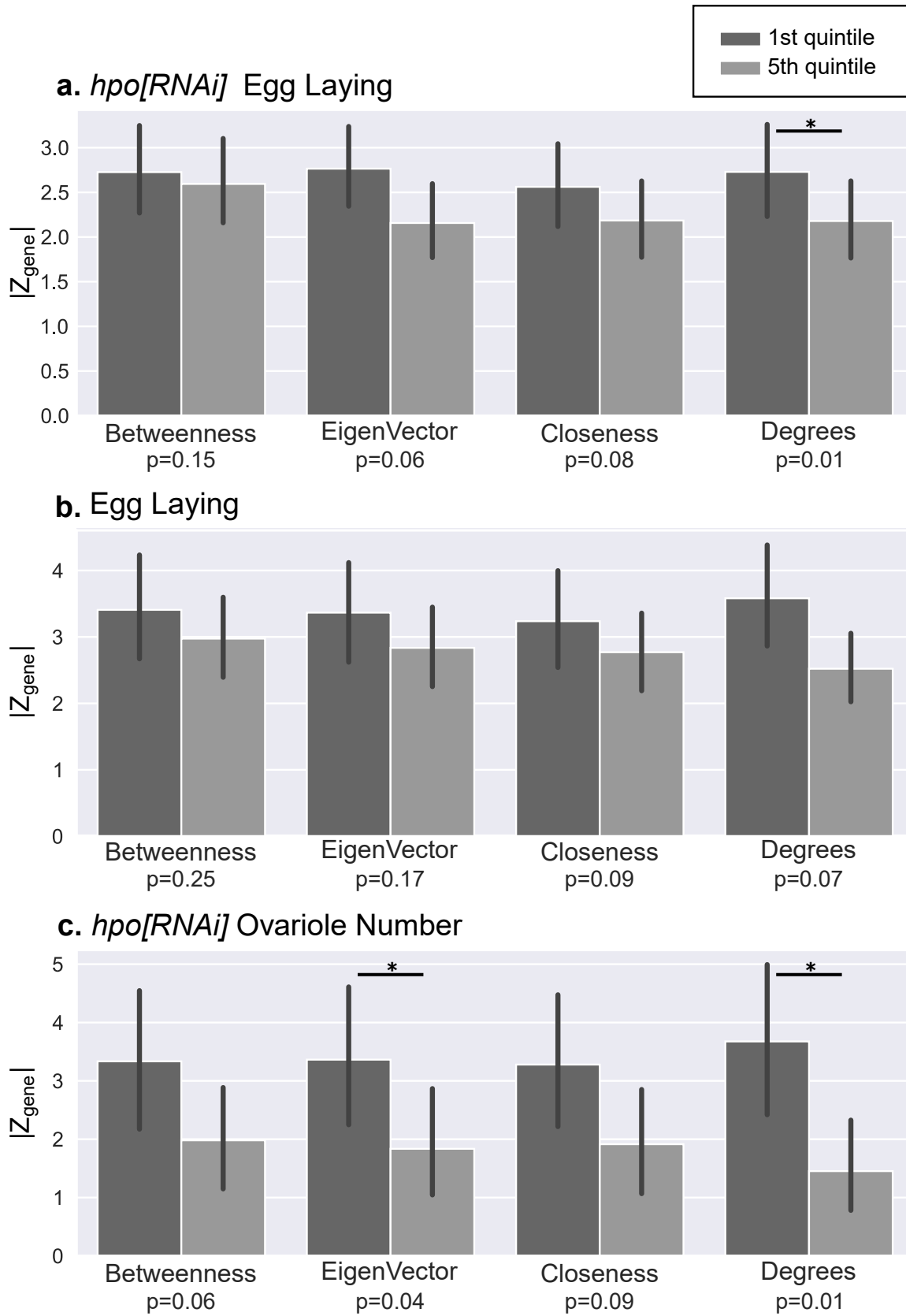


Figure C.4 (following page): Comparisons of the Z_{gene} scores of the positive candidate genes sorted by centrality metrics. In each screen (a, b, c), the $|Z_{gene}|$ values of the first (dark grey) and fifth (light grey) quintiles of positive candidate genes ordered by rank for each of the four chosen centrality metrics, are plotted as a bar plot. Bars indicate standard error, and significant differences (p-value<0.05 Mann Whitney U test) are indicated by asterisks. p-values are displayed below every bar plot.

Figure C.4: (continued)



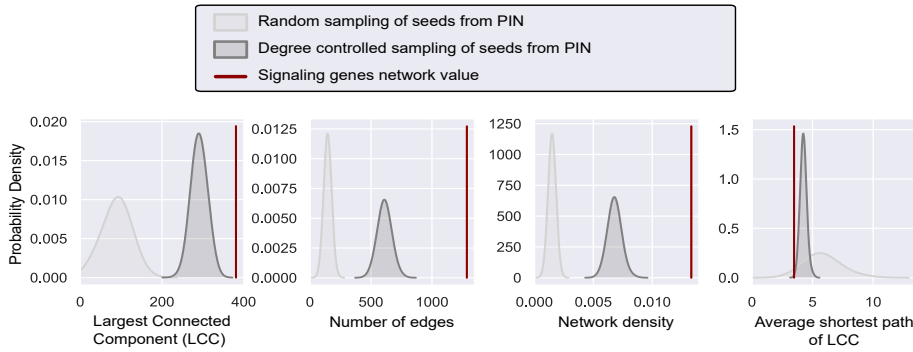
Sub-Network	Number of Seeds	Number of Connectors	Number of connector genes above $ Z_{gene} $ threshold within sub-network phenotype
<i>hpo[RNAi]</i> Egg Laying	58	18	7 (41.1%)
Core	27	10	1 (10.0%)
Egg Laying	49	11	0 (0.0%)
<i>hpo[RNAi]</i> Ovariole Number	66	11	3 (27.2%)

Table C.3: Distribution of seed genes and connectors in each sub-network. Two genes that were above $|Z_{gene}|$ threshold (Table 2 and Figure 7- Figure supplement 2) in the *hpo[RNAi]* Egg Laying (CG12147) and *hpo[RNAi]* Ovariole Number seed list (CG6104) were not found in the PIN, and therefore not included in the network analysis or in this table (see methods for details). The removal of these two genes accounts for the difference between the number of positive candidates in Table 2 and Figure 7- Figure supplement 2, and the number of seed genes in these two sub-networks (Supplementary File 1 and Figure 7- Figure supplement 1). The proportion of connectors whose loss of function produced a significant phenotype ($|Z_{gene}|$ above threshold) is in parentheses and plotted in Figure 7a, 7b). All connectors except *eukaryotic translation initiation factor 3 subunit j (eIF3J)* in the *hpo[RNAi]* Egg Laying sub-network, for which no RNAi line was available at the time of testing, were tested. Therefore, the percentages of connectors above the threshold for the *hpo[RNAi]* Egg Laying sub-network were calculated out of 17 connectors.

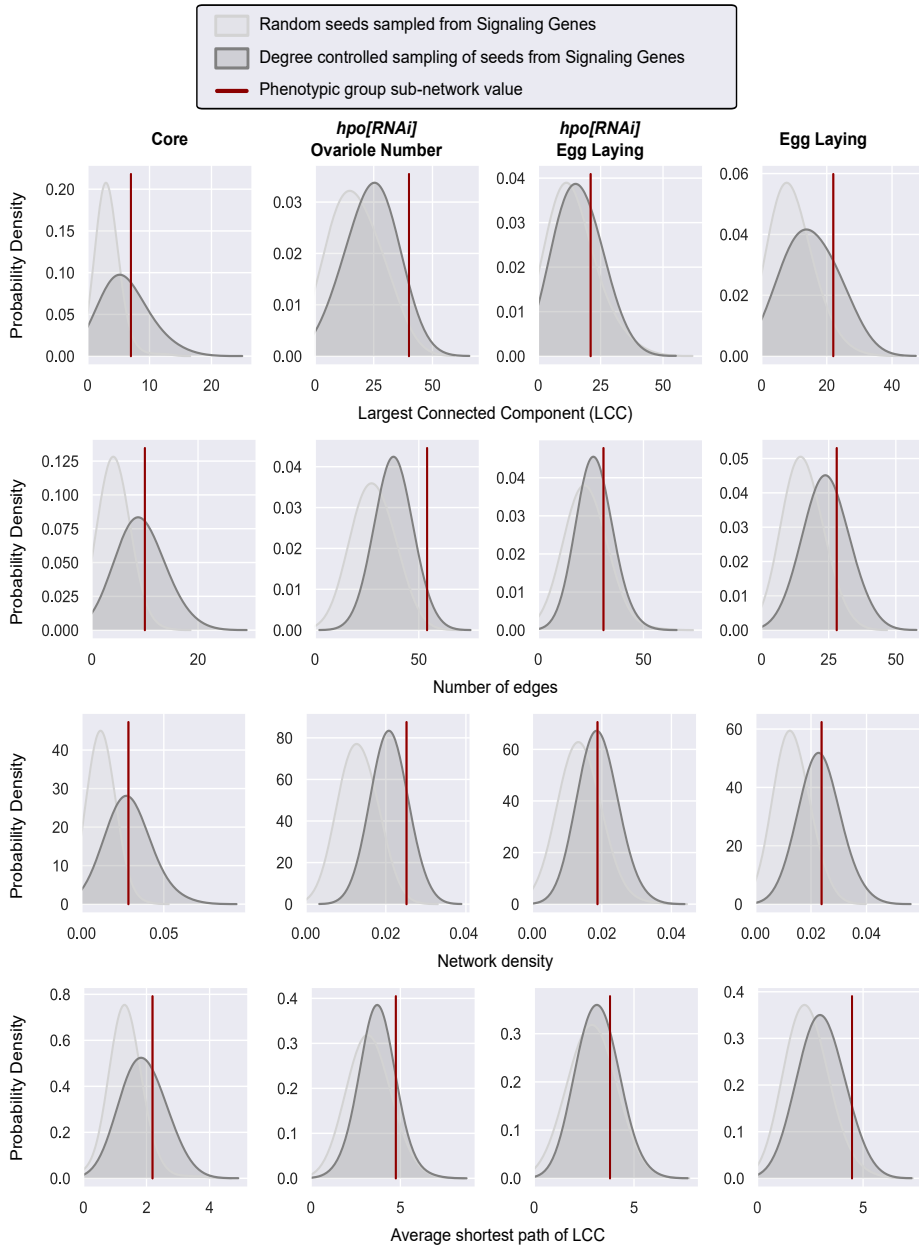
Figure C.5 (following page): Comparison of network metrics of seed lists obtained from the screen. **a)** Comparison of network metrics of all screened genes (red line) to two null distributions of network metrics derived by randomly sampling an equal number of random genes (light grey curve) or degree-controlled genes (dark grey curve) from the PIN. **b)** Comparisons of the Largest Connected Component (LCC), network density, number of edges and average shortest path between the seed network (red line) and the randomly sampled null distribution (1000 random samples) of both an equal number of random genes (light grey curve) or degree-controlled genes (dark grey curve) from the PIN.

Figure C.5: (continued)

a. Screened gene background



b. Phenotypic sub-networks



hpo[RNAi] Egg Laying Module

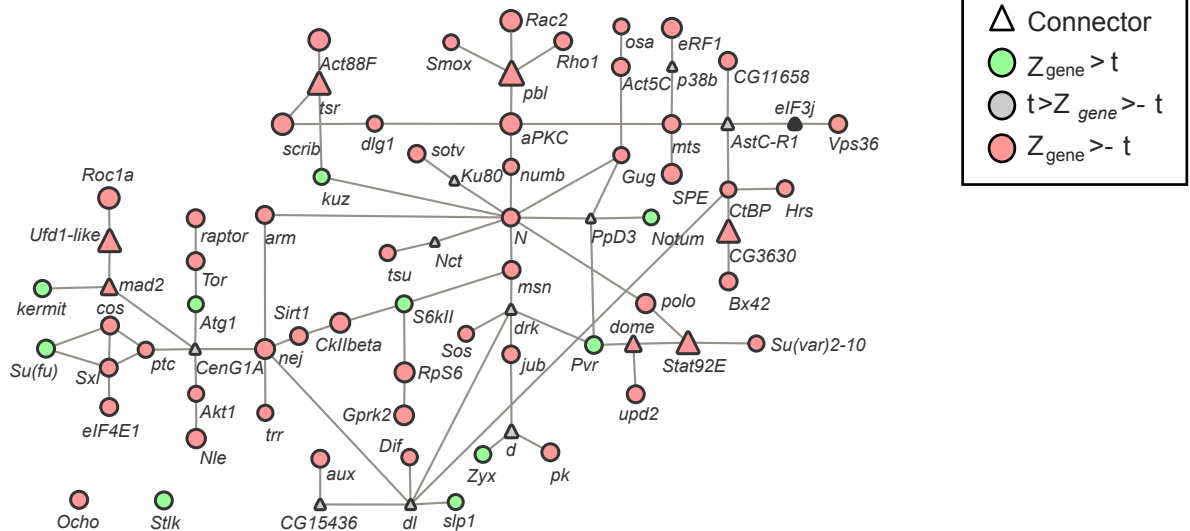


Figure C.6: *Hpo*[RNAi] Egg Laying Sub-Network generated by the Seed Connector Algorithm (SCA). The size of the shapes indicates the Z_{gene} score of the gene. Circles = seed genes; triangles = connector genes. Green = genes with a positive Z_{gene} above the threshold. Red = genes with a negative Z_{gene} above threshold. Grey = genes with Z_{gene} values below the threshold. All connectors were phenotypically tested (Supplementary File 1) except *eukaryotic translation initiation factor 3 subunit j* (*eIF3J*), in the *hpo*[RNAi] Egg Laying Module (black triangle), for which no RNAi stock as available at the time of testing.

Egg Laying Module

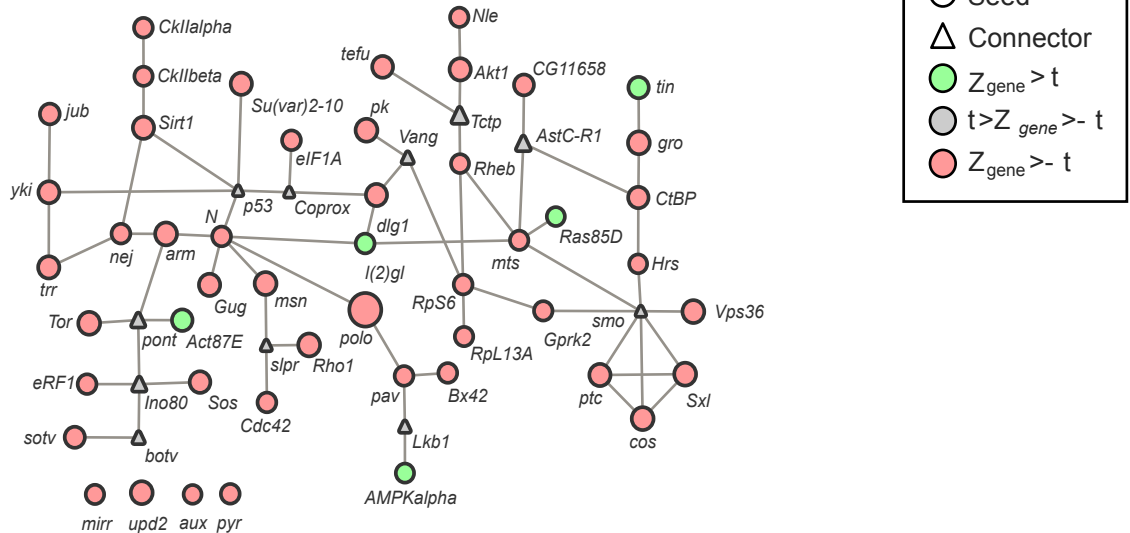


Figure C.7: Egg Laying Sub-Network generated by the Seed Connector Algorithm (SCA). The size of the shapes indicates the Z_{gene} score of the gene. Circles = seed genes; triangles = connector genes. Green = genes with a positive Z_{gene} above the threshold. Red = genes with a negative Z_{gene} above threshold. Grey = genes with Z_{gene} values below the threshold. All connectors were phenotypically tested (Table C.1).

hpo[RNAi] Ovariole Number Module

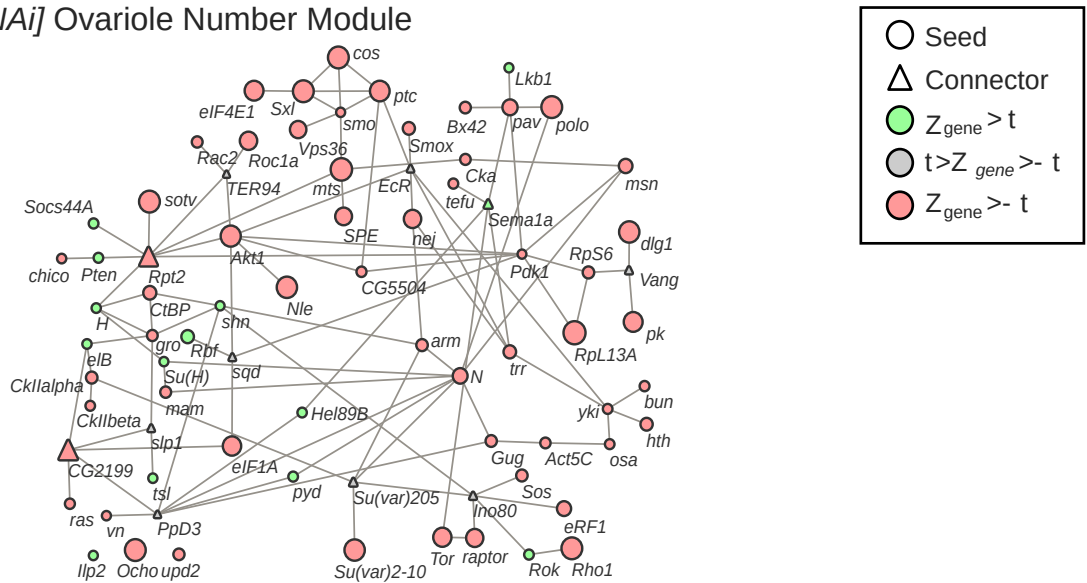


Figure C.8: *Hpo[RNAi]* Ovariole Number Sub-Network generated by the Seed Connector Algorithm (SCA). The size of the shapes indicates the Z_{gene} score of the gene. Circles = seed genes; triangles = connector genes. Green = genes with a positive Z_{gene} above the threshold. Red = genes with a negative Z_{gene} above threshold. Grey = genes with Z_{gene} values below the threshold. All connectors were phenotypically tested (Supplementary File 1).

Unique Genes with RNAis		Number of genes above threshold in				
		<i>hpo[RNAi]</i> Egg Laying $Z_{gene}> 1 $	Egg Laying $Z_{gene}> 5 $	<i>hpo[RNAi]</i> Egg Laying $Z_{gene}> 5 $	<i>hpo[RNAi]</i> Ovariole Number $Z_{gene}> 2 $	All three screens
Connectors	42	32	10/32 (31.2%)	13/42 (30.9%)	12/32 (37.5%)	8/32 (25%)
Signalling candidates	463	273	49/273 (17.9%)	59/463 (12.7%)	67/273 (24.5%)	27/273 (9.8%)

Table C.4: Number of unique connector and signalling genes above $|Z_{gene}|$ threshold for the three phenotypes measured in this screen (Figure 1a, 1b, 1c). Table indicates the number of unique genes among the connector genes and signalling genes screened, that had available RNAi lines at the time of analysis. The number of genes in the *hpo[RNAi]* Egg Laying sub-network that were above the primary filter of $|Z_{gene}| > 1$ are also indicated. Percentage of the number of connectors and signalling candidates above threshold for each phenotype from the number of connectors above the primary filter is in parentheses and plotted in Figure 7c. All connectors except *eukaryotic translation initiation factor 3 subunit j (eIF3J)* in the *hpo[RNAi]* Egg Laying sub-network, for which no RNAi line was available at the time of testing, were tested. Therefore, the percentages of connectors above the threshold were calculated out of 32 unique connectors.

a. Network modules predicted by SCA

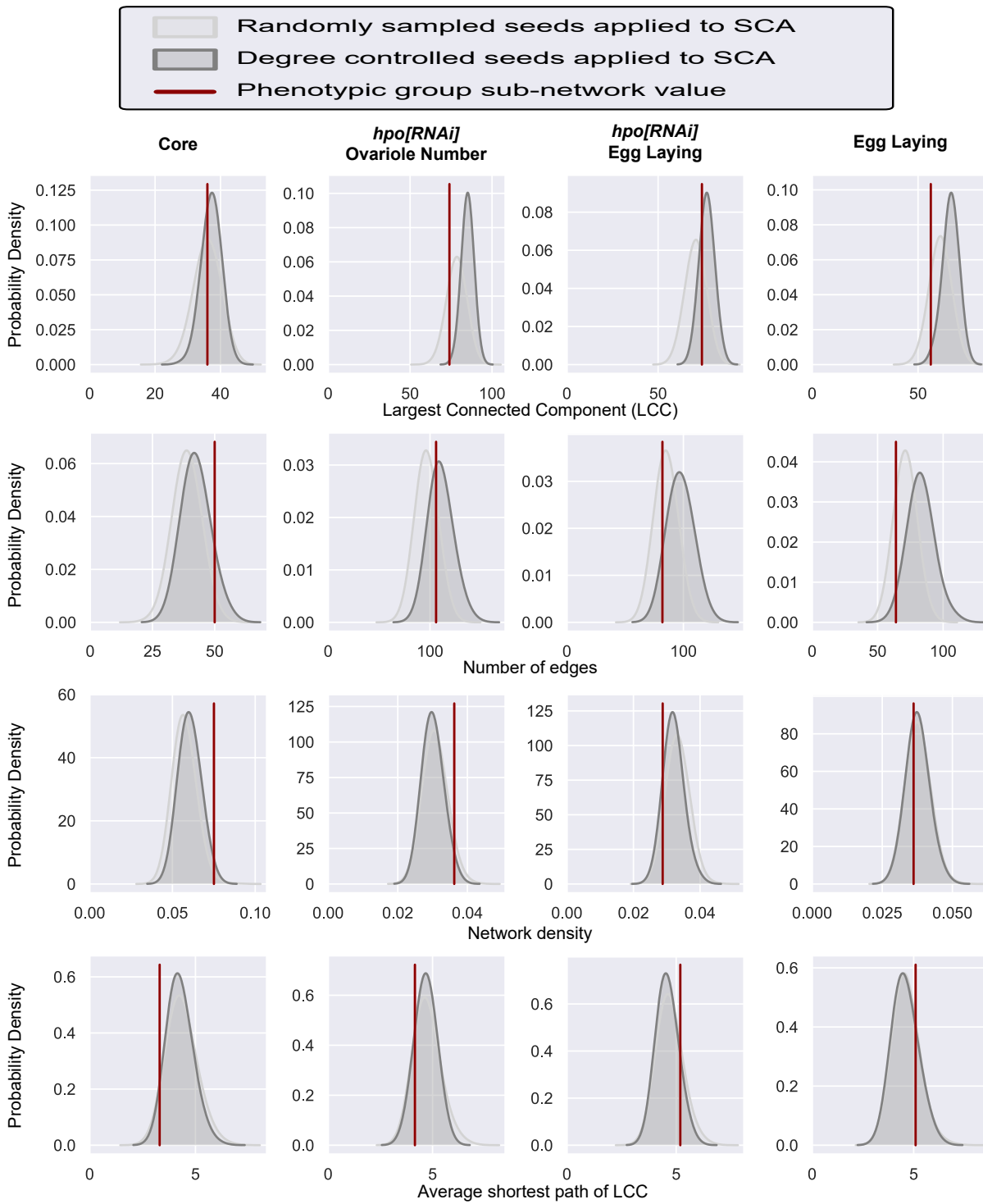


Figure C.9: Centrality metrics of the sub-networks. Box plots of the four centrality measures calculated for the genes in each of the four phenotypic putative modules generated by the SCA, shown in Figures 5 and Figures 5- Figure supplements 1,2 and 3.

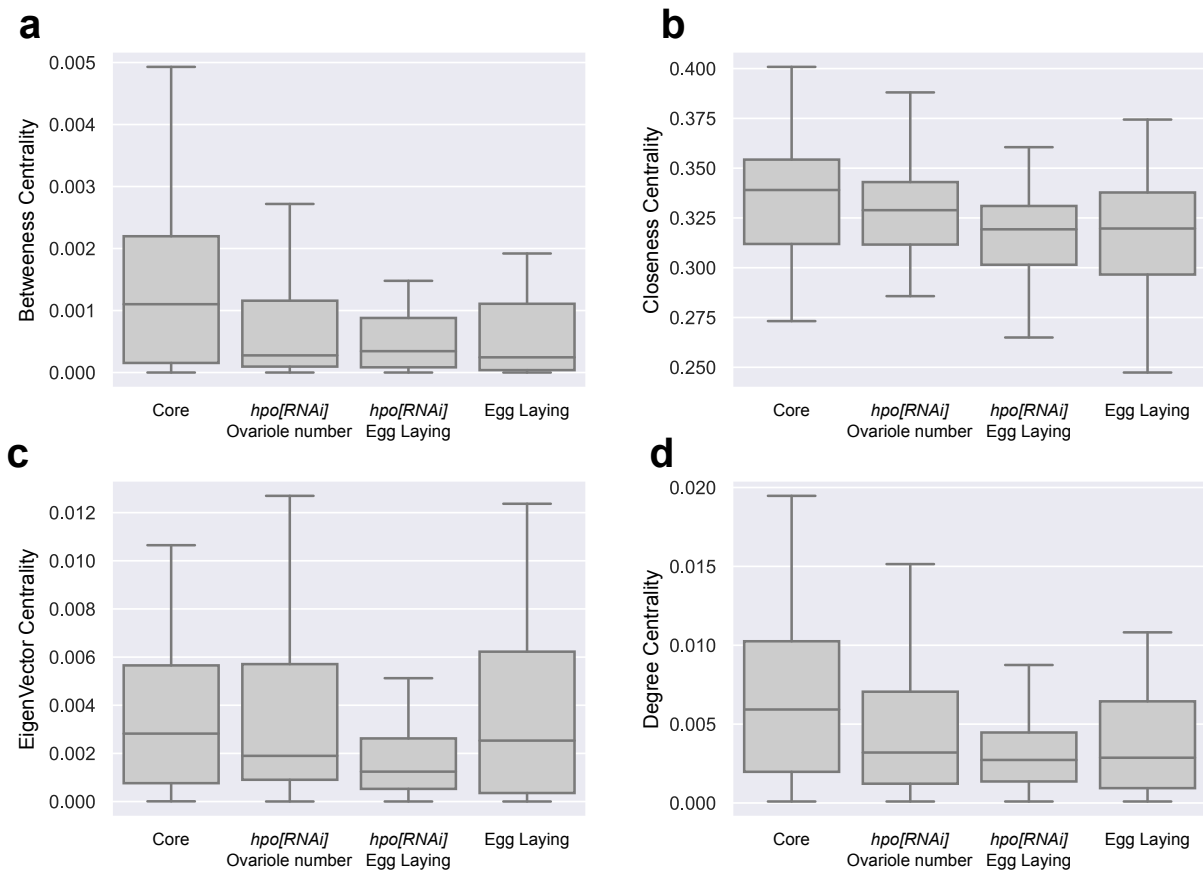


Figure C.10: Comparison of network metrics after application of the seed connector algorithm (SCA). a) Comparisons of the Largest Connected Component (LCC), network density, number of edges and average shortest path of the sub-networks obtained by applying the SCA to a list of seed genes. The red lines indicate the network metrics of the sub-network (Figure 5b and Figure 5- Figure supplement 1-3) obtained by applying the SCA to the four seed lists based on the outcomes of the phenotypic screens. The curves display the null distributions of each network metric for the sub-networks. These distributions were obtained by applying the SCA to 1000 seed gene lists randomly selected from among the signalling genes tested in our functional screen (light grey curve), or from a degree-controlled seed list (dark grey curve).

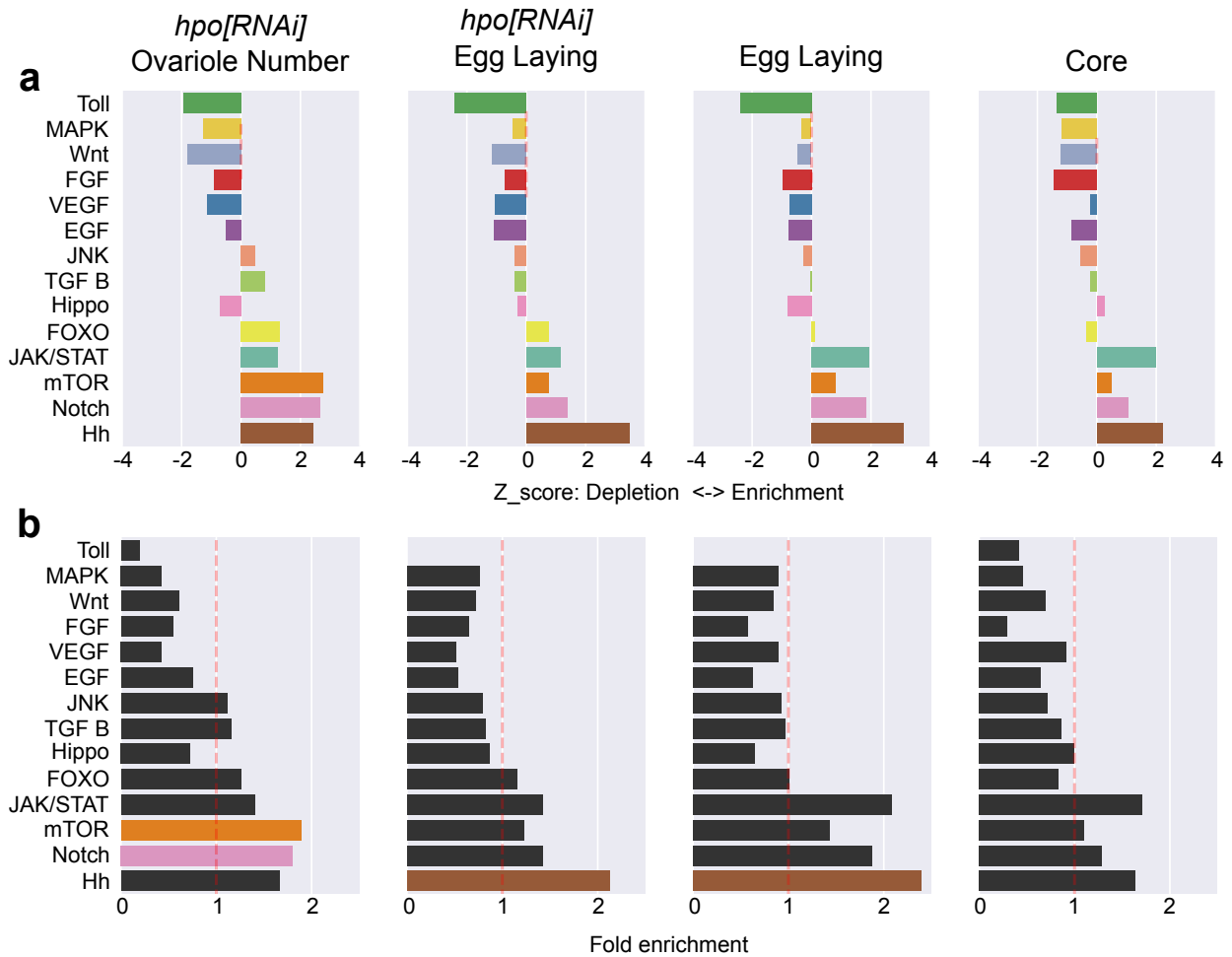


Figure C.11: Signalling pathway enrichment/depletion analysis. a) For each putative module generated by the SCA, a null distribution of the expected number of members of a signalling pathway from a group of the same number of randomly selected signalling pathway genes was calculated. The Z score from the expected distribution was then calculated. Negative Z scores represent a depletion, while positive Z scores represent an enrichment. No single pathway is enriched in any of those SCA-generated putative modules. b) Fold enrichment and hypergeometric p-value calculation for each pathway in the four SCA-generated putative modules. Pathway members in colour (orange = mTor; brown = Hedgehog; pink = Notch) have a p-value < 0.05.

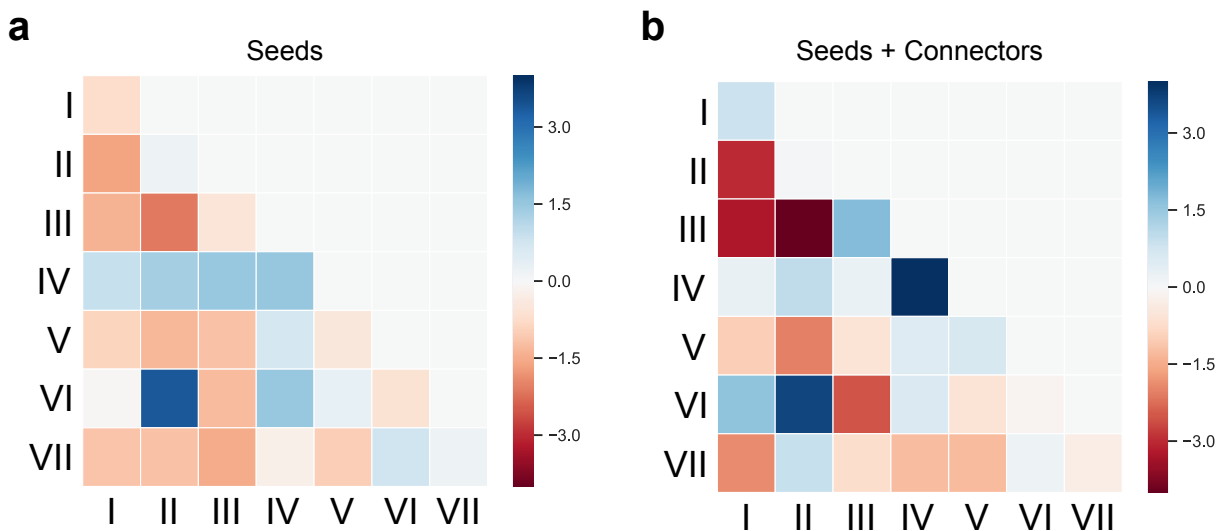


Figure C.12: Comparison of edge densities between the seven sub-networks of the meta network, and to a randomly assigned grouping of genes in the meta network. a) Z scores of the value of the edge densities of only the seed genes in the seven groups of the meta network, compared to the distribution of edge densities of a random assignment of the same seed genes to seven sub-networks of the same size (1000 repeats). **b)** Z scores of the value of the edge densities of both the seed and connector genes in the seven sub-groups of the meta network, compared to the distribution of edge densities of a random assignment of the same genes to seven sub-networks of the same size (1000 repeats). Blue: meta network sub-network has higher edge density than the random group; red: meta network sub-network has lower edge density than the random group.



THIS THESIS WAS TYPESET using \LaTeX , originally developed by Leslie Lamport and based on Donald Knuth's \TeX . The body text is set in 11 point Source Sans Pro. I modified the Dissertate template that can be used to format a PhD thesis with this look and feel. It has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.

Thank you for reading this thesis until the end.