

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Detection, Quantification, and Characterization of Diverse RNA Transcript Isoforms

Permalink

<https://escholarship.org/uc/item/63v3493d>

Author

Hateley, Shannon Layne

Publication Date

2016

Peer reviewed|Thesis/dissertation

Detection, Quantification, and Characterization of Diverse RNA Transcript Isoforms

by

Shannon Layne Hateley

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lior Pachter, Co-Chair
Professor Michael Eisen, Co-Chair
Professor Jennifer Doudna
Professor Martyn Smith

Fall 2016

Abstract

Detection, Quantification, and Characterization of Diverse RNA Transcript Isoforms

by

Shannon Layne Hateley

Doctor of Philosophy in Molecular and Cell Biology

and the

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Lior Pachter, Co-Chair
Professor Michael Eisen, Co-Chair

RNA sequencing (RNA-Seq) is a powerful technique to assess gene expression and to characterize RNA transcripts, as it allows for high-throughput study of RNA at a nucleotide resolution. It can provide transcriptome-wide information on alternative splicing, novel transcripts, tissue-specific and allele-specific expression, and RNA-DNA differences. While high-throughput sequencing has revolutionized RNA research, there are several challenges associated with its application. The limiting factors in the utility of any RNA-Seq experiment are two-fold: (1) at the experimental stage, correctly isolating and preparing the sample so as to capture all transcripts of interest, and (2) at the data analysis stage, utilizing the right computational methods to accurately assign sequencing reads to the correct transcripts. Non-canonical isoforms can easily be missed by standard methods at either stage. As RNA exists in a multitude of forms corresponding to its diverse functions, with variations in sequence, length, shape, and number, it is important to develop sequencing methods that are capable of capturing this diversity.

Herein, we describe methods for performing RNA sequencing experiments and data analysis, illustrate their utility in experimental studies using *Drosophila melanogaster*, and present novel methods for the detection, quantification, and characterization of two non-canonical transcript isoforms: the fusion transcript and circular RNA.

*To Mom,
for ardently encouraging my curiosity.*

Table of Contents

Abstract	I
Table of Contents	ii
List of Figures	iii
List of Tables	iv
Acknowledgements	v
1 Introduction to RNA	I
The Complexity of RNA	2
Preparation and Sequencing of RNA	4
RNA-Seq Data Analysis	5
Overview of Dissertation	7
2 Transcriptomic Response of <i>Drosophila melanogaster</i> Pupae Developed in Hypergravity	9
1. Introduction	10
2. Results and Discussion	11
2.1 <i>Hypergravity Accelerates Time to Pupal Development</i>	11
2.2 <i>Hypergravity Induces Consistent Transcriptional Response</i>	12
2.3 <i>Extrapolation with qRT-PCR</i>	15
2.4 <i>Major Processes Involved in Hypergravity Response</i>	16
2.5 <i>Additional Transcriptional Response in Hypergravity</i>	24
2.6 <i>Hypergravity Versus Pro-Oxidant-Induced Transcriptional Response</i>	25
3. Conclusions	26
4. Materials and Methods	26
3 Fast Fusion Detection, Assembly, and Quantification from RNA Sequencing Data	30
1. Introduction	31
2. Methods	33
2.1 <i>kallisto Fusion Stage</i>	33
2.2 <i>pizzly Fusion Stage</i>	34
3. Performance Assessment	36
3.1 <i>Datasets</i>	37
3.2 <i>Tools for Benchmarking</i>	37
3.3 <i>Results</i>	38
4. Future Steps	40
5. Supporting Information	42
4 Circular RNA in <i>Drosophila melanogaster</i> embryos	43
1. Introduction	44
2. Results and Discussion	46
2.1 <i>A Test of Circular RNA Sample Purification Methods</i>	46
2.2 <i>Staged Single Embryo Collection and Sequencing</i>	48
2.3 <i>Computational Workflow for Circular RNA Detection</i>	49
3. Future Directions	51
Bibliography	53
Appendix	62

List of Figures

Figure 1.1	Timeline of RNA discoveries	3
Figure 1.2	Overview of standard RNA-Seq experiment	6
Figure 1.3	Overview of typical RNA-Seq workflow	7
Figure 2.1	Comparison of sex-specific genes to ascertain bias in data	14
Figure 2.2	Hypergravity-induced gene expression and functional categories	15
Figure 2.3	Comparative gene expression from RNA-Seq and qRT-PCR	16
Figure 2.4	Hypergravity-induced changes in transport genes	20
Figure 2.5	Hypergravity affects defense genes	22
Figure 2.6	Hypergravity alters chitin biosynthesis and cuticular genes	23
Figure 3.1	Fusion genes resulting from chromosomal rearrangements	31
Figure 3.2	Reads meeting specific criteria are passed to pizzly	34
Figure 3.3	A check for alignment to partner transcript ensures correct fusion	35
Figure 3.4	A read can be split-aligned between the two fusion transcripts	36
Figure 3.5	Split alignment search	36
Figure 4.1	Circular RNA research timeline	45
Figure 4.2	Division of reads from samples treated with RNase or rRNA depletion	47
Figure 4.3	<i>D. melanogaster</i> embryos at six developmental stages	48
Figure 4.4	Principal component analysis plot of five embryo stages	49
Figure 4.5	Circular RNA fragment mapping in RNA-Seq	50
Figure 4.6	Workflow for circRNA mapping and transcript assignment	51
Figure A.1	Cluster map of hypergravity genes versus stress response genes	62

List of Tables

Table 2.1	Count of pupae reaching three developmental stages at different gravities I I	
Table 2.2	Hypergravity-induced changes in redox genes	18
Table 2.3	Read details for samples	28
Table 3.1	Fusion detection tools compared to pizzly for benchmarking	38
Table 3.2	75nt positive control	39
Table 3.3	100nt positive control	39
Table 3.4	100nt negative control	39
Table 3.5	Spike-in real dataset containing nine true synthetic fusions	40
Table 3.6	Original kallisto read counts for suspected fusion transcripts	40
Table 3.7	Kallisto is able to correctly assign reads to fusions at the transcript level	41
Table 3.8	Kallisto reported TPM for the nine synthetic fusion transcripts	41
Table 4.1	Ribosomal reads remaining post treatment	47
Table A.1	Positive PANTHER overrepresentation test for biological processes	63
Table A.2	Positive PANTHER overrepresentation test for molecular function	65
Table A.3	PANTHER enrichment test for biological processes	67
Table A.4	PANTHER enrichment test for molecular function	68
Table A.5	Gene expression data (qRT-PCR) at 3 g, 5 g and 8 g	69
Table A.6	Software details	70
Table A.7	qRT-PCR primers	71

Acknowledgements

It is commonly accepted that grad student life is demanding, stressful, lonely, and all too often, depressing. Many an article, PhD comic, and blog have articulated the grueling nature of this highest pursuit. It is with all the more wonder then that I find myself here, at the end of my PhD, rather unscathed. How am I still just about as bright-eyed and bushy-tailed as a first-year? The answer is: you guys. I wholeheartedly owe my very deepest acknowledgements to the people around me who have given so much support, inspiration, and love at all the moments when I was most in need.

First and foremost I thank my advisor, Lior Pachter, who gave me both freedom and guidance throughout my graduate career. His knowledge, mentorship, sense of ethics, (and sense of humor) draws the best people to his group. I cannot express how fortunate I feel to have been in the Pachter lab. My labmates are truly amazing. Thank you Lorian Schaeffer for being there on the daily and commiserating with me in all things wetlab, Harold Pimentel for saving my data and the servers more times than I can count, Isaac Joseph for letting me work on fusions and for motivating me to finally learn linear algebra. Shannon McCurdy, Bo Li, Nicolas Bray, Adam Roberts, Brielin Brown, Atif Rahman, Lynn Yi, Faraz Tavakoli, Meromit Singer, Robert Tunney, Sharon Aviron, Mathilde Paris, and Natth Bejraburnin - thank you for the many conversations, both academic and otherwise.

I owe much gratitude to Maren Bell, who helped me with the daunting task of setting up an entire laboratory and who is one of the kindest and most practical listeners I've ever known.

I thank my collaborators who contributed their knowledge and expertise to this work: at NASA, Ravikumar Hosamani and Sharmila Bhattacharya, and at University of Iceland, Páll Melsted. I thank my committee for guidance on this dissertation. I thank the Eisen lab for help with all things *Drosophila*, especially Steven Kuntz, Carolyn Elya, and Jackie Villalta. To the flies - thank you for your lives. I thank my many mentors in MCB and Comp Bio for imparting so much wisdom over the years. And I am grateful for my classmates and all the fun we had together.

Outside of school, my friends, Hermes, Karta, my dance family, and the spirit of Berkeley itself nourished my soul and fed my artistic and adventurous side.

I must particularly thank my great friend Jennifer van Velkinburgh and my mentor Callum Bell, without whose encouragement I probably never would have thought it possible to even go to grad school. And I thank the extraordinary people at St. John's College, especially Howard Fisher, who inspired me to be a scientist. Of course I thank the science, the math family I join, and the countless work hours and countless shoulders I stand upon.

Lastly, I thank my family for being there all throughout grad school and always, and who love me limitlessly, whether I'm a doctor or a bum.

I Introduction to RNA

The Complexity of RNA

For every defined rule in molecular biology there are discovered as many specific exceptions to that rule. Take, for example the following:

Rule: The genetic code corresponds to twenty amino acids and three stop codons.

Exception: Selenocysteine is conditionally coded via the umber stop codon in the presence of selenium and a characteristic stem-loop structure sequence (Böck et al. 1991).
And
Variations on the standard code exist for human mitochondrial, *Candida albicans*, and *Mycoplasma* genes (Barrell, Bankier, and Drouin 1979; Santos and Tuite 1995; Yamao et al. 1985).

Rule: DNA exists in a right-handed double helical structure.

Exception: Left-handed Z-form DNA exists (Mitsui et al. 1970).

Rule: Translation begins at the 5' end of mRNA molecules.

Exception: Internal ribosome entry sites (IRES) allow for translation in an end-independent manner (Pelletier and Sonenberg 1988).

Rule: Proteins are the only biological molecules that can catalyze chemical reactions.

Exception: Ribozymes are RNA catalysts (Kruger et al. 1982) and it is the ribosome that performs protein synthesis (Nissen et al. 2000).

New discoveries regularly reshape our understanding of the molecular world. In many cases a finding that at first appears to be an exception to the rule, with further study grows into its very own branch of research. This phenomenon is no more prominent than in the study of RNA.

Originally defined as a support structure and passive intermediary between DNA and protein, RNA has been found to participate in everything from epigenetic regulation to enzymatic activity. Small RNAs can act as scaffolds in chromatin structure modification and transcriptional silencing (Holoch and Moazed 2015), are a central cofactor in RNA splicing, and can guide DNA methylation (Morris and Mattick 2014). Catalytic RNAs, or ribozymes, perform protein synthesis and cleave self- and other- RNAs (Walter and Engelke 2002). RNA can be modified *in vivo* from the DNA sequence from which it was transcribed either with RNA-editing (even restoring the ancestral sequence from a less-stable DNA mutant (Bar-Yaacov et al. 2013)), or with 3'-5' splicing, abnegating the linearity of DNA sequence by becoming a circularized transcript (Salzman et al. 2012). The vast agency of RNA grows more apparent with each characterization of a new RNA class. Figure 1.1 presents an abridged timeline of RNA discovery, illustrating the diversity RNA roles. By most of today's estimates, the number of non-coding genes in humans is greater than the

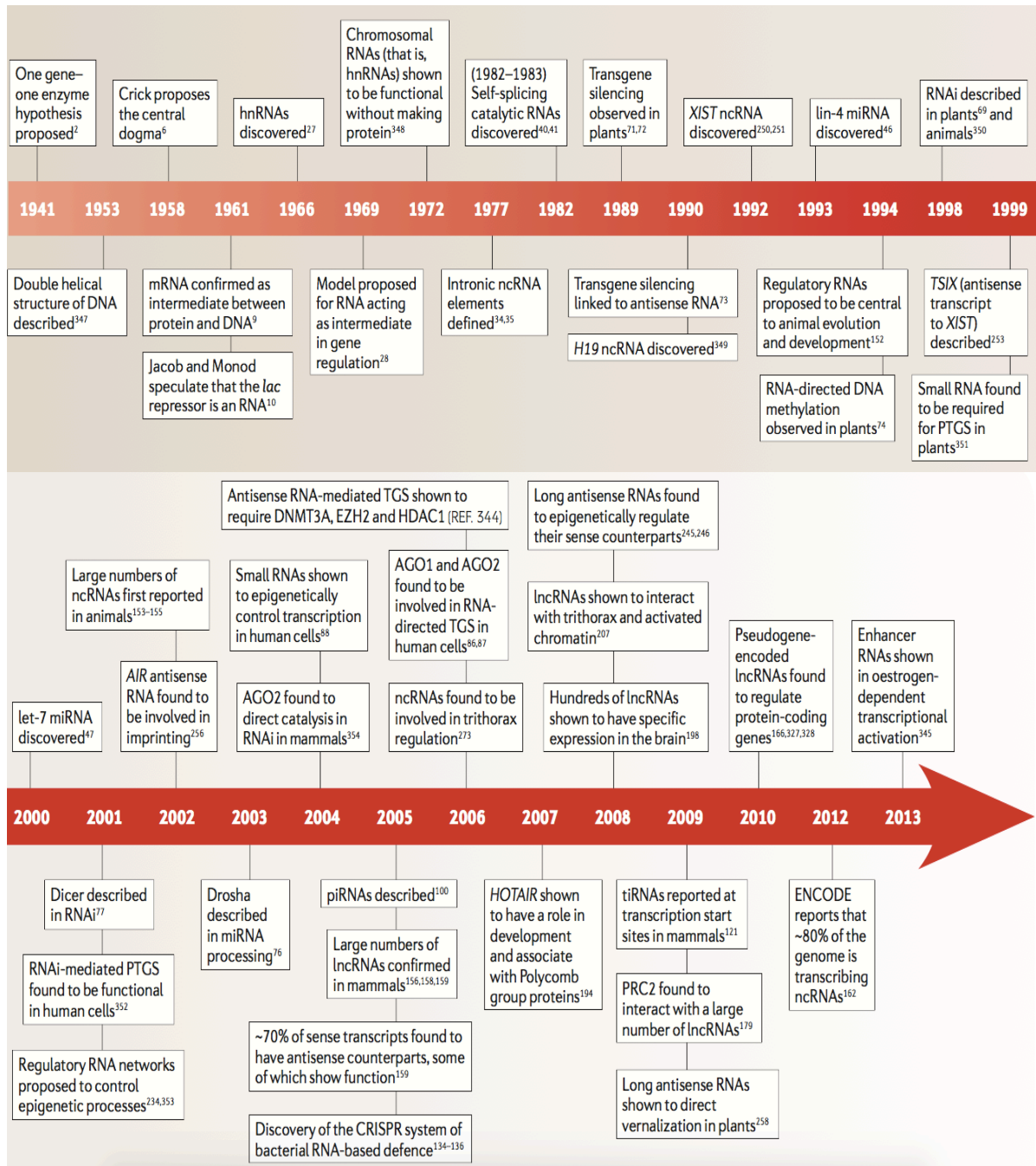


Figure 1.1 Timeline of RNA discoveries

Reproduced from Morris and Mattick 2014. Abbreviations: AGO, Argonaute; *AIR*, also known as *AIRN* (antisense of *IGF2R* non-protein coding RNA); CRISPR, clustered regularly interspaced short palindromic repeat; DNMT3A, DNA (cytosine-5)-methyltransferase 3A; ENCODE, Encyclopedia of DNA Elements; EZH2, enhancer of Zeste 2; *H19*, H19 imprinted maternally expressed transcript; HDAC1, histone deacetylase 1; hnRNA, heterogeneous nuclear RNA; *HOTAIR*, HOX transcript antisense RNA; lncRNA, long non-coding RNA; miRNA, microRNA; ncRNA, non-coding RNA; piRNA, PIWI-interacting RNA; PRC2, Polycomb repressive complex 2; PTGS, post-transcriptional gene silencing; RNAi, RNA interference; TGS, transcriptional gene silencing; tiRNA, transcription initiation RNA; *XIST*, X inactive specific transcript.

number of protein coding genes, suggesting that the focus in multicellular organisms should be shifted away from gene-level analysis and towards transcriptional analysis, regarding “genetic loci as ‘fuzzy’ transcription clusters” (Morris and Mattick 2014). The emerging view of the transcriptional landscape is one of dynamic, interconnected complexity. RNA has come a long way from mere messenger: it takes part in transcription, translation, and protein function. RNA is the Swiss Army knife of the cellular world.

Preparation and Sequencing of RNA

Because of its central role in so many aspects of cellular function, discovering and characterizing RNA in its many forms is essential to understanding molecular and cell biology. Much of current RNA research takes advantage of advances in sequencing technology. RNA sequencing (RNA-Seq) facilitates fast, sensitive, and high-throughput study of RNA at single-nucleotide resolution (Han et al. 2015). Paired with downstream analysis methods, RNA-Seq can identify new and existing RNA transcripts, characterize sequence anomalies such as differential splicing and sequence variants, and quantify transcript abundance. Multiple platforms have been developed for RNA-Seq. This dissertation focuses on the most popular platform, Illumina sequencing-by-synthesis (Mortazavi et al. 2008). In a standard RNA-Seq experiment, as outlined in Figure 1.2, purified RNA molecules are converted into a library of template DNA suitable for high-throughput sequencing through a series of molecular transformations. First, the RNA is fragmented using divalent cations and temperature elevation. Next, reverse transcriptase and random primers copy the RNA fragments into first-strand complementary DNA (cDNA). DNA polymerase synthesizes the second strand of the cDNA. The double-stranded cDNA undergoes end repair, adenylation, and adapter ligation. The prepared fragments are purified and enriched using PCR to create the final cDNA template library. This library is submitted to cluster generation and finally to sequencing on the sequencing-by-synthesis machine.

While high-throughput sequencing has been revolutionary in the detection and characterization of RNA isoforms, the procedure is not without its challenges. Ribosomal and transfer RNA make up approximately 80% and 15% of total RNA, respectively (Lodish et al. 2000). The interest of most RNA-Seq experiments is in the remaining 5% of the sample. Sequencing RNA *en toto* would dilute the sequencing of any transcripts of significance. When performing RNA-Seq for most applications, it is important to reduce the amount of ribosomal RNA present in the sample. The standard method for RNA-Seq library preparation in eukaryotes utilizes poly(A⁺) selection, a method to retain only those transcripts which are polyadenylated, namely messenger RNA (mRNA). This method is commonly referred to as mRNA-Seq and allows for analysis of protein-coding gene expression, gene fusion detection, and alternative splicing. For analysis of non-coding and other non-polyadenylated RNAs, ribosomal RNA depletion of total RNA is performed before sequencing preparation. This method is becoming more common as the study of non-canonical, non-coding

RNA becomes more popular and ribosomal depletion becomes more efficient. Many RNA selection, depletion, and library preparation techniques have been developed for isolating various types of transcripts including methods for small RNA, targeted regions, and direct RNA sequencing (Ozsolak et al. 2009). Awareness of the biases and trade-offs in sequencing library preparation, and alteration of methods according to the desired outcome, enables researchers to use RNA-Seq to detect a multitude of RNA transcript isoforms.

RNA-Seq Data Analysis

Advances in experimental methods have increased the power of RNA-Seq for transcript detection. Improvement of the downstream computational steps is equally important. RNA-Seq data analysis is an active field of research, composed of multiple steps: processing the RNA reads, providing accurate annotations, quantifying transcript abundance, determining gene expression across samples, inferring systems-level interactions, and generating meaningful data visualizations. See Figure 1.3 for a typical RNA-Seq workflow. While much work has been done in the field, there is still much room for improvement. The utility of high-throughput sequencing has been limited by the bioinformatic challenges associated with processing and analyzing the vast amount of output data (Conesa et al. 2016). Many RNA-Seq analysis programs require significant computational resources, time, and computer science and statistics expertise, all of which limit the utility of such programs to biologists. As RNA-Seq continues to increase in popularity because of its ability to study transcriptome-wide expression in a vast array of experiments, establishing faster and more robust methods for RNA-Seq data analysis will be of great benefit to the research community.

In standard RNA-Seq data analysis, reads are aligned to a reference genome or transcriptome and transcript abundance is estimated. This can be a lengthy process. For example, using the common TopHat2-Cufflinks workflow (Kim et al. 2013, 2; Trapnell et al. 2012) on 20 samples of 30 million RNA-Seq reads each, alignment takes 28 core hours on 20 cores and quantification takes another 14 hours. Newer methods that utilize streaming algorithms (Roberts and Pachter 2013) or k-mer matching (Patro, Mount, and Kingsford 2014) have been introduced that speed up the processing, albeit with less accurate transcript quantification. Most recently, the application of k-mer pseudoalignment (Bray et al. 2016b) has sped up RNA-Seq quantification by two orders of magnitude from previous leading programs, while also improving abundance estimation. Utilizing this fast, computationally inexpensive method for isoform detection allows for iterative development of best-in-class detection tools, an example of which is described in detail in Chapter Three.

Most RNA-Seq analysis workflows have been designed for detection and quantification of reads from linear mRNA transcripts. As non-coding RNA and non-canonical isoforms appear to make up the majority of RNA expression, there is an

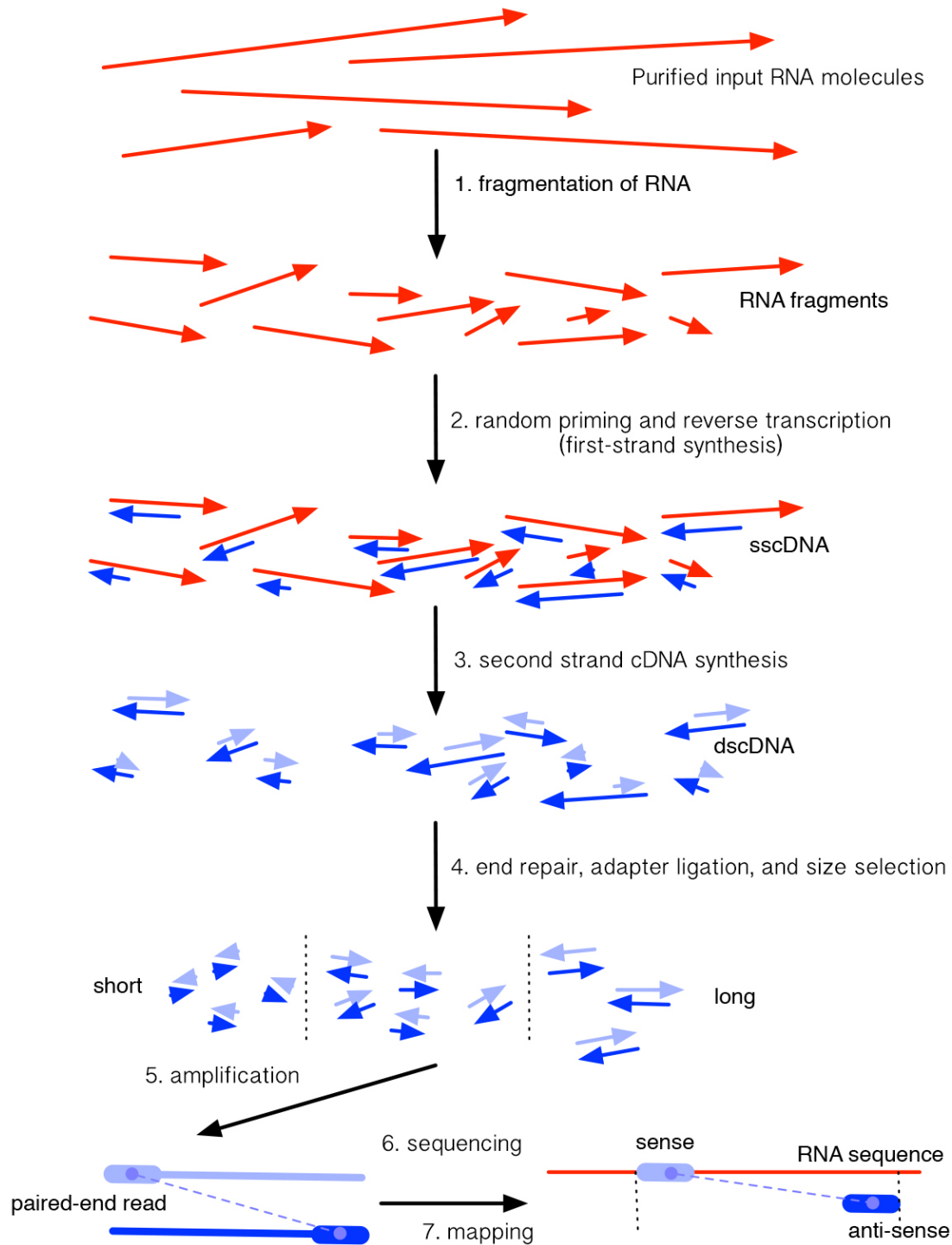


Figure 1.2 Overview of standard RNA-Seq experiment

Purified RNA (via ribosomal depleted or poly(A⁺) selection), is fragmented (1) followed by first strand synthesis using random priming and reverse transcription (2). After second strand synthesis (3), double-stranded cDNA is subject to end-repair, adapter ligation, size selection (4), and amplification (5), resulting in fragments, around 300nt, suitable for sequencing (6). Sequenced reads are mapped to opposite strands of the genome (7) or to a reference transcriptome. All arrows are directed 5' to 3' in transcript orientation.

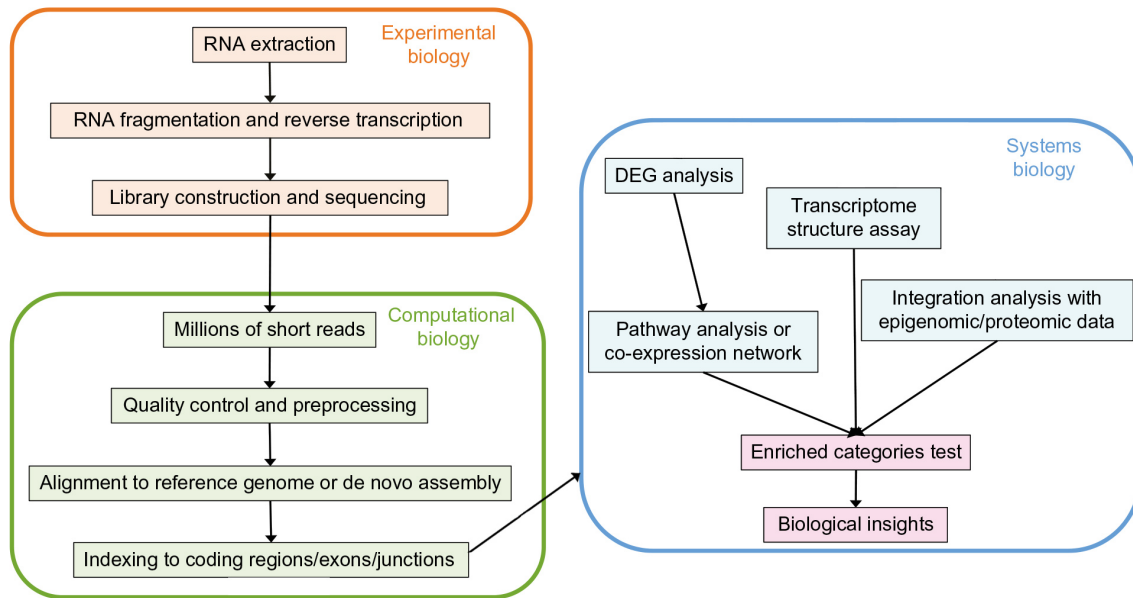


Figure 1.3 Overview of typical RNA-Seq workflow

Reproduced from Han et al. 2015. RNA-Seq can be divided into three parts: experimental wet-lab work, computational biology of processing reads and assigning reads, and biological interpretation.

obvious need to shift focus from solely characterizing linear, polyadenylated mRNA, to a more diverse detection strategy. In order to detect non-canonical isoforms, the standard rules must once again be redefined.

Overview of Dissertation

RNA functions in a multitude of ways, and its characterization is a costly and challenging process. Therefore, it is vital to develop methods to recognize transcript isoforms that otherwise remain unobserved. Herein, we describe methods for performing RNA sequencing and analysis, illustrate their utility in experimental studies, present examples of non-canonical RNA transcript isoforms, and suggest novel methods for their detection.

Chapter Two presents techniques to create, analyze, and compare datasets from research conducted in collaboration with the Biomedical Performance and Behavior Laboratory in the Space Biosciences Division of NASA. Focusing specifically on one of many studies conducted in this collaboration, it presents a systems-level study of *D. melanogaster* pupae developed in hypergravity. RNA sequencing lends itself exceptionally well to the exploratory nature of space research, as it can answer a wide range of questions and offer strategies to probe into many aspects of functional and molecular biology. Efficient sample processing methodologies, coupled with tools for downstream analysis, are particularly important here because the scarce resources and

technical constraints of spaceflight often limit the number of studies possible. Improved RNA-Seq methodologies will provide a robust, relevant resource for standard lab-based studies, space biology research, and beyond.

Chapter Three presents the development and testing of an RNA-Seq analysis tool for detecting a non-canonical form of transcript: the fusion transcript. Through reanalysis of “unaligned” reads, the detection, assembly, and quantification of fusion transcripts is achieved. This method exemplifies the canon that new discoveries are often waiting to be found in previously discarded data.

Chapter Four presents detection of circular RNA across time points in developing *D. melanogaster* embryos utilizing single embryo collection, specialized molecular processing, and modification of standard downstream computational tools. This illustrates the full workflow of an RNA sequencing experiment, from experimental design, biological sample generation, RNA extraction and library creation, analysis design, and transcript detection and quantification.

2 Transcriptomic Response of *Drosophila melanogaster* Pupae Developed in Hypergravity

This chapter has been previously published as an open access article, with reprint approval by all authors, as:

Hateley, S., Hosamani, R., Bhardwaj, S. R., Pachter, L. & Bhattacharya, S. Transcriptomic response of *Drosophila melanogaster* pupae developed in hypergravity. *Genomics* 108, 158–167 (2016).

I. Introduction

During the pupal stage, most holometabolous insects undergo an extensive metamorphosis from larva into winged adult. Larval structures are degenerated while adult structures are formed from the imaginal discs and histoblasts. This metamorphosis is tightly regulated transcriptionally, and is evolutionarily conserved. It also depends on Earth's gravity. For example, the orientation of the pupa is dependent upon gravitational cues for proper alignment (Jander 1963) and deviation to microgravity or hypergravity environments can profoundly influence metamorphosis and alter associated gene expression (Herranz et al. 2010). Comprehensive gene expression analysis is therefore necessary to understand the underlying molecular response of developing pupae in altered gravity environments.

To date, gene expression studies on the response to altered gravity have mostly utilized microarray technology. Herranz et al. measured transcriptional response to simulated micro and hypergravity in *D. melanogaster* pupae via microarray, discovering that a significantly higher number of genes related to metamorphosis were changed in hypergravity (2 g) than in simulated microgravity (Herranz et al. 2013). Pupae exposed to even higher gravitational forces (6 g and 12 g) exhibited altered mRNA expression across multiple pathways, upsetting diverse physiological functions and behaviors (Herranz et al. 2013). In another microarray study by Herranz et al., imagoes developed entirely in a diamagnetic-levitation-induced microgravity or simulated hypergravity (2 g) environment revealed changes in gene expression linked to immunity, stress, and temperature-response (Herranz et al. 2012). Other research efforts, focused on the effect of hypergravity on aging, physiological stress, and behavior (Le Bourg et al. 2000), have shown that chronic hypergravity (3 g and 5 g for two weeks) has a beneficial effect on longevity, resistance to heat, and behavioral aging in adult *D. melanogaster*. Conversely, hypergravity in combination with cold stress can have negative impact on these same traits (Le Bourg 2012). Lifelong exposure of flies to different hypergravity (4 g, 5 g, 7.38 g) conditions can also have negative effects on longevity (Minois 2006). The wide-ranging scope of these results highlights a complex interplay between biology and gravity. Further characterization of the effects of altered gravity on an organism would benefit from the use of RNA sequencing, with its ability to address global transcriptional changes with reliable accuracy, higher sensitivity in measurement, and power to detect novel transcripts.

Here, we have characterized the effects of hypergravity on the *D. melanogaster* transcriptome via differential expression analysis of RNA-Seq data from 3 g-exposed and 1 g-control pupal samples. Our data reveal significant alterations in the expression of genes involved in ion transport, oxidative stress, immune function, proteolysis, and cuticle development in response to chronic hypergravity in *D.*

melanogaster. Understanding the role of gravity in development contributes to knowledge of basic developmental biology and is a critical aspect of space biology research.

2. Results and Discussion

2.1 Hypergravity Accelerates Time to Pupal Development

To see if gravity-induced gene expression changed in proportion to increase in gravitational force, in addition to 3 g samples, we also allowed *D. melanogaster* to develop at 5 g and 8 g for evaluation using qRT-PCR. Pupae that developed at 3 g from egg to stage P6 (Campos-Ortega and Hartenstein 1985) did not show any change in developmental phenotype. However, as we increased gravitational force from 3 g to 5 g, pupal development accelerated significantly, as evidenced by a higher number of advanced stage pupae and increased eclosion rate compared to 3 g (Table 2.1). Next, we increased g-force to 8 g and looked for developmental defects. At 8 g, egg laying and hatching were minimal. While at 1 g and 3 g around 150 pupae developed, at 8 g only 16 were counted. Among the 16, 9 developed into 3rd instar larvae and died before pupation, and only 7 attained pupal stage. The 8 g environment was detrimental to normal development. The low count of pupae and adults at 8 g (Table 2.1) illustrate that the force exceeded the viability threshold. Those that did manage to pupate developed at a much faster rate than at 5 g. Previous studies have shown that heat stress can accelerate growth in *Drosophila* (Petavy et al. 2001). To ascertain whether this phenotype was due to a change in gravity or due to heat stress, we fixed an iButton to the rotating arm inside the centrifuge to monitor the temperature profile through the course of the experiment. The iButton recorded the temperature every 15 minutes until the end of the experiment. The temperature ranged from 22.5 to 24.5 °C inside the centrifuge, indicating no significant change in both the 3 g and 5 g environments. These data rule out the possibility of heat stress influencing the developmental process in pupae. Thus, chronic hypergravity significantly influenced pupal development.

Table 2.1 Count of pupae reaching three developmental stages at different gravities

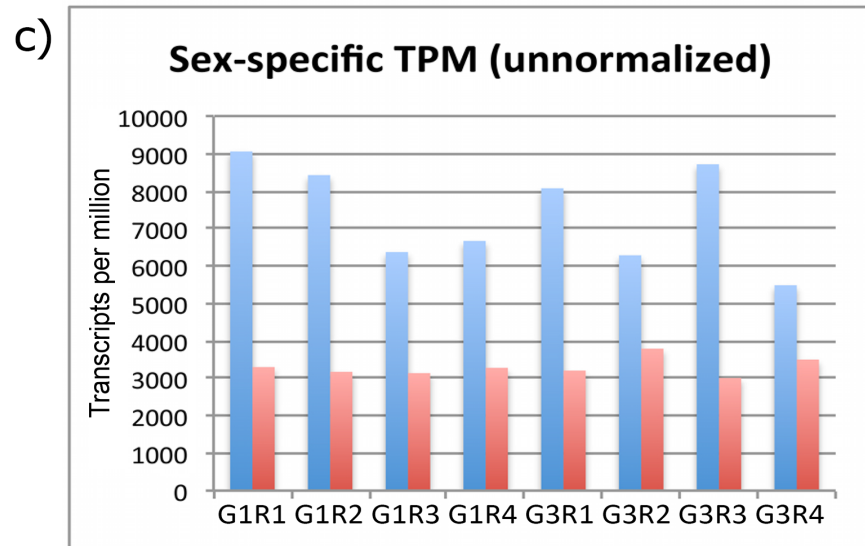
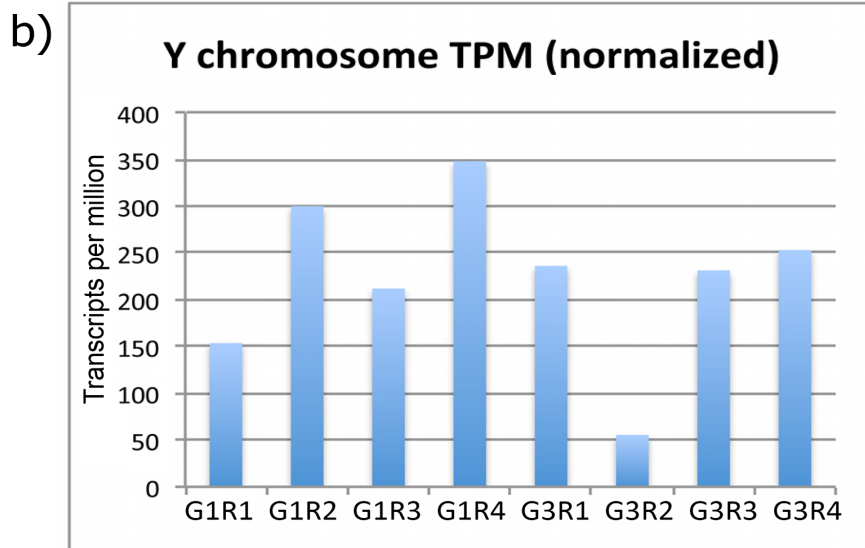
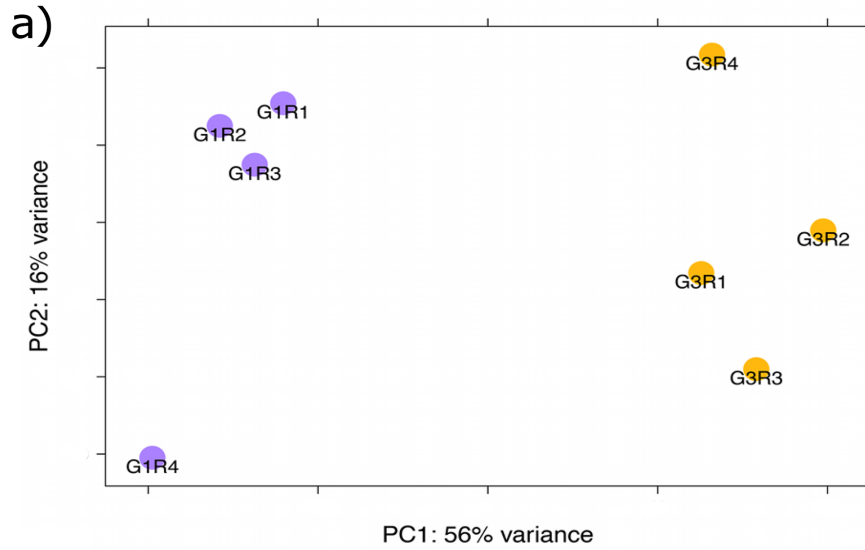
	1 g	3 g	5 g	8 g
Early pupae	151	149	142	0
Dark pupae	2	3	16	7
Adult eclosion	1	1	8	0

Pupal development was monitored and samples were counted immediately post-centrifugation. While there was no change in pupal development at 1 g and 3 g, at 5 g we observed some acceleration in pupal development. At 8 g there was a profound effect on development and survival as was evident from the low pupal counts.

A clear change in developmental phenotype was evident as we increased the gravitational acceleration from 5 g to 8 g. However, we were specifically interested in understanding the transcriptional response to 3 g for two reasons. First, during space flight, astronauts are not only exposed to microgravity but also experience around 3 g during launch and 3 g or more during reentry. More importantly, chronic hypergravity models have been exploited to complement and predict microgravity-associated changes (Herranz et al. 2013). Since 3 g exerts enough force for chronic hypergravity exposure without decreasing overall viability, we sought to elucidate 3 g-induced transcriptional changes in *D. melanogaster* pupae in our study.

2.2 Hypergravity Induces Consistent Transcriptional Response

To discern the transcriptional response to hypergravity, we performed differential expression analysis on control and 3 g-exposed pupae. First, poly(A+) libraries from four control and four 3 g samples were pooled and sequenced on one lane of the Illumina HiSeq 2000, yielding a total of 214.5 million Casava quality-filtered reads (Bolger, Lohse, and Usadel 2014). After trimming, reads were aligned to the *D. melanogaster* reference transcriptome using Bowtie 2 (Langmead and Salzberg 2012) and transcript abundances estimated using eXpress (Roberts and Pachter 2013). Principal component analysis (PCA) revealed the gene expression data are well clustered between different biological replicates of both control and hypergravity-exposed samples, suggesting hypergravity is the greatest variable factor across conditions (Figure 2.1a). Due to technical constraints (see Section 4.2 for details), pupae could not be sexed before sample preparation. To check for possible sex bias of pupae and ensure that sex was not a confounding factor in our analyses, we compared Y chromosome and sex-specific gene expression across samples. Y chromosome transcripts confirm that males were present in each sample, with expression patterns uncorrelated to PCA components 1 and 2 (Figure 2.1b). Comparison to the sex-specific transcripts in identically-staged pupae as detailed by Lebo (Lebo et al. 2009) show only 19 of the same genes to be significantly differentially expressed, of which there is no sex-specific bias per sample or between conditions (Figure 2.1c,d). Differential expression analysis with DESeq2 (Love, Huber, and Anders 2014) revealed 1,513 genes were significantly ($q < 0.05$) altered in response to chronic hypergravity, of which 704 were upregulated and 809 were downregulated in 3 g samples compared to control (For complete table of genes, with functional classifications, see Supplemental Table S1 of published edition (Hateley et al. 2016) of study). Gene ontology analysis was performed using the PANTHER classification system (Mi et al. 2016). Among the 1,513 genes differentially expressed, as many as 475 genes have no defined information on their molecular functions or biological processes. This study focuses upon the remaining 1,038 genes. Based upon the hierarchical relation of gene ontology molecular and biological associations (Gene Ontology Consortium 2016), we divided the 1,038 genes into 13 functional



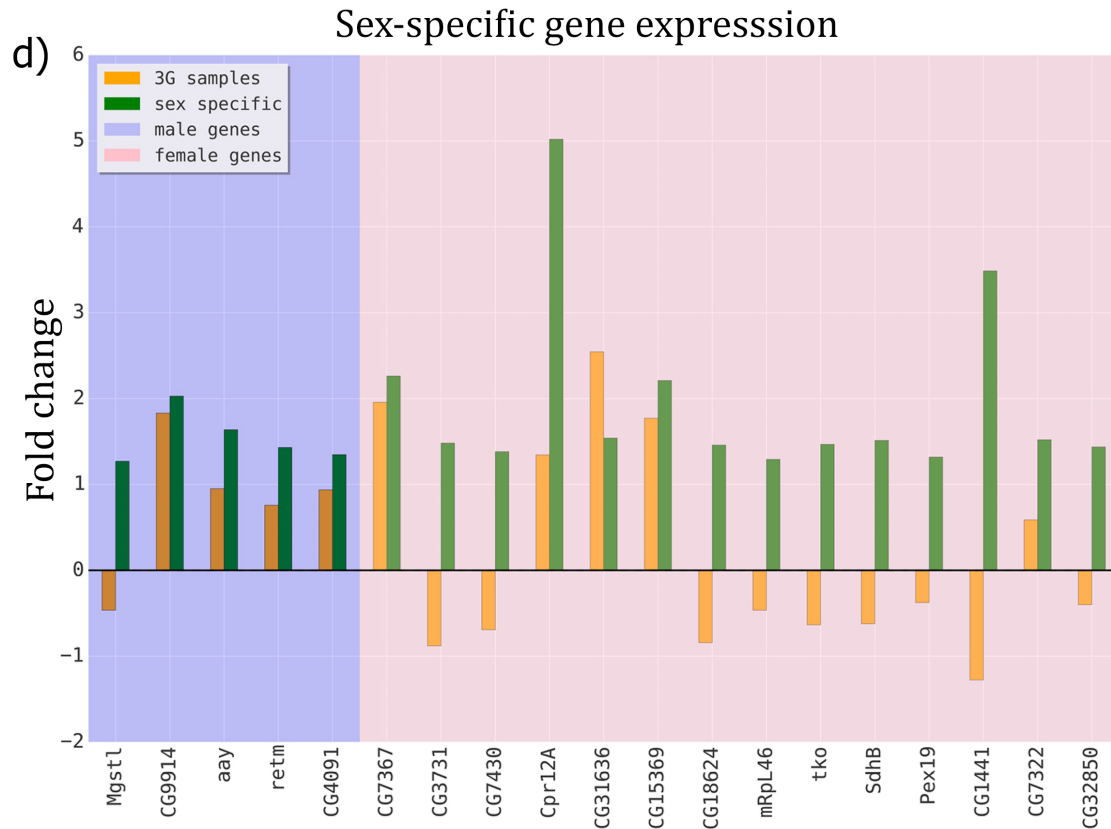


Figure 2.1 Comparison of sex-specific genes to ascertain bias in data

Previous page: a) Principal component analysis (PCA) shows samples segregate along PC1 (principal component 1) between treatment (gold) and control (purple). b) Y chromosome-specific gene expression, normalized for across-sample comparison, shows at least one male was included in each pooled sample. The variation in Y-originating transcript TPM (transcripts per million) does not correspond to PC1 or PC2 in (a), indicating sex is not a major variable factor between conditions. c) Expression of the 19 genes from (d) shows per-sample ratio of male-specific (blue bars) to female-specific (red bars) genes do not correlate with principal components 1 and 2. This page: d) Fold-change expression of sex-specific genes in male (blue background) and female (red background) P6 pupae (green bars) as detailed in Lebo et al. compared to log2 fold-change expression in this study's hypergravity pupae (gold bars). Spearman correlation is not significant (Male genes (n = 5): $\rho = 0.900$, $p\text{-val} = 0.037386$, female genes (n = 14): $\rho = 0.3231$, $p\text{-val} = 0.259874$), indicating sex of pupae is not a major confounding factor between 1 g control and 3 g treatment samples. G1R1 = 1 g force control, replicate 1. G3R1 = 3 g force treatment, replicate 1.

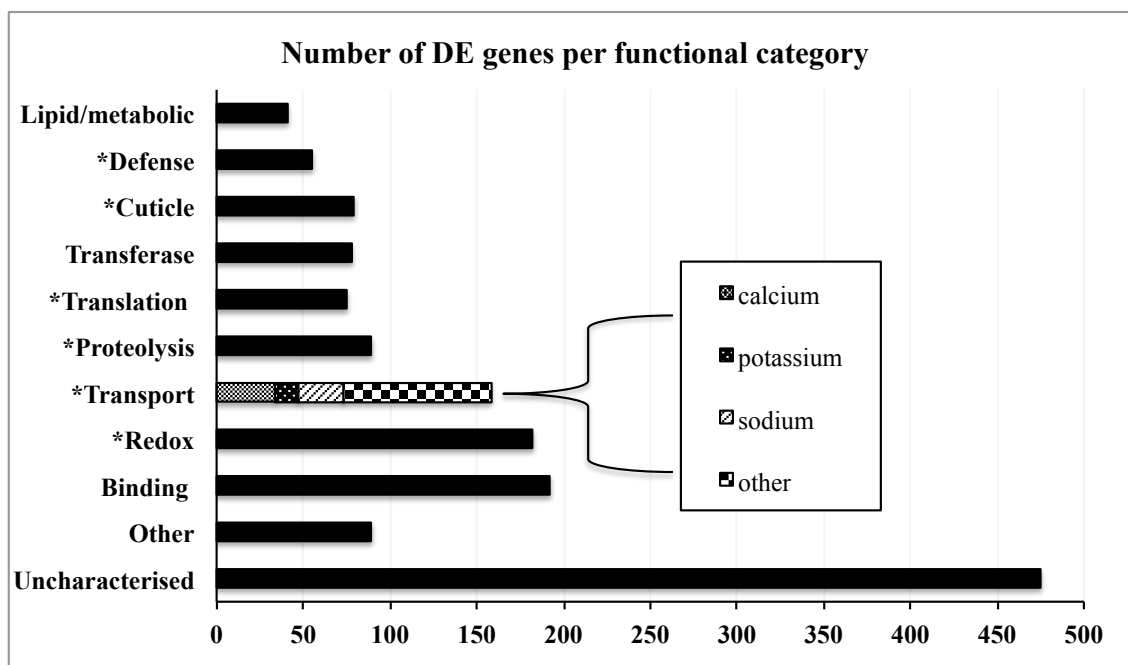


Figure 2.2 Hypergravity-induced gene expression and functional categories

Based on gene ontology analysis, the differentially expressed ($q \leq 0.05$) genes in pupae developed at 3 g were assigned to 13 different categories. This classification was guided by the molecular function of the genes and their associated biological processes. Significant categories as determined via statistical overrepresentation and enrichment tests are indicated with asterisks.

classifications (Figure 2.2). Using PANTHER statistical overrepresentation and enrichment tests along with these functional classifications, we determined that the 3 g treatment predominantly affected five major processes: ATP synthesis/redox, ion channels/transport, defense/immunity, cuticle development, and proteolysis. Overrepresentation tests reported enrichment of ion channel/transport genes up to 5.3 times ($\text{padj} = 5.70\text{E-}05$), cuticular genes up to 3.0 times ($\text{padj} = 1.68\text{E-}05$), and redox genes up to 6.0 times ($\text{padj} = 4.37\text{E-}04$) their expected levels compared to control genes. Translation related genes were also overrepresented up to 2.2 times ($\text{padj} = 7.95\text{E-}06$), but with minimal fold change in expression. The enrichment tests found defense and proteolysis genes to be enriched ($\text{padj} = 2.27\text{E-}03$ and $3.08\text{E-}08$, respectively), and redox genes to be under-enriched ($\text{padj} = 6.59\text{E-}03$). (For results of statistical tests, see Table A.1-A.4). The major affected processes are discussed in detail in Section 2.4.

2.3 Extrapolation with *qRT-PCR*

We selected 18 differentially expressed genes representing the diverse biological functions for further study using quantitative RT-PCR. Comparison of RNA fold-change and pattern of gene expression between RNA-Seq data and quantitative RT-

PCR is highly consistent across biological replicates (Figure 2.3) with Spearman correlation = 0.885, $p = 1.075e-06$.

After confirming our RNA-Seq results from the 1 g and 3 g conditions, we examined whether there was any relationship between changes in gene expression and changing gravitational force by performing qRT-PCR on pupae exposed to increasing gravity levels (3 g, 5 g and 8 g). There was no clear dose-responsive pattern observed in gene expression of these genes at the higher g-levels (see Section 4.5 for methods and Table A.5 for qRT-PCR data). This is consistent with the fact that higher g-loads seem to induce additional stressful effects on the flies including affecting development at 5 g and loss of viability at 8 g (Table 2.1).

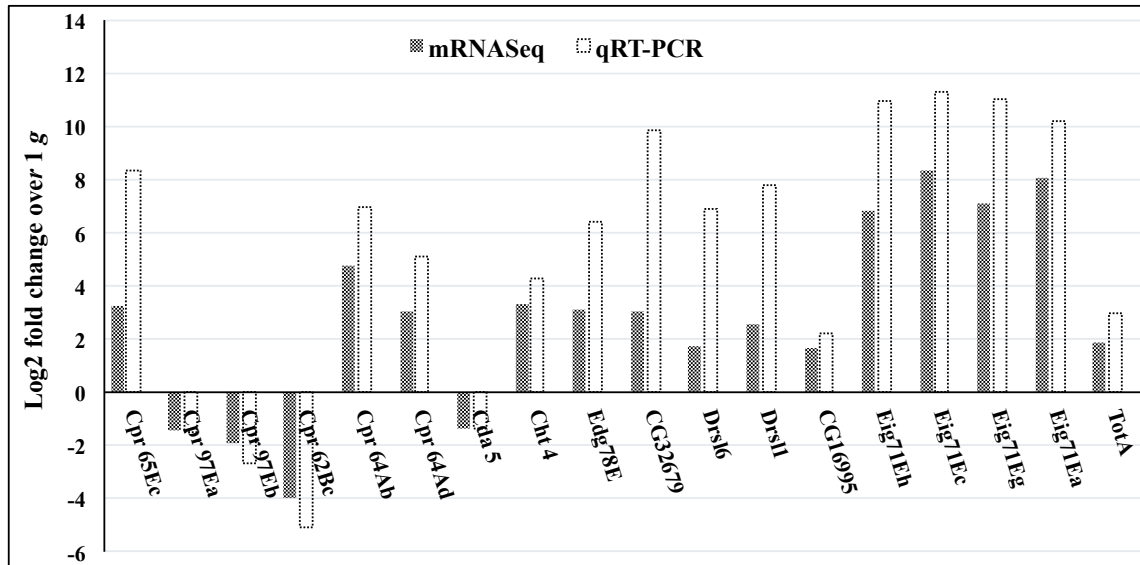


Figure 2.3 Comparative gene expression from RNA-Seq and qRT-PCR

Comparative log2 fold change expression of representative genes across the diverse categories of functions obtained by RNA-Seq and qRT-PCR analysis.

2.4 Major Processes Involved in Hypergravity Response

2.4.1 Hypergravity Alters Expression of Genes Encoding Mitochondrial Electron Transport, Oxidative Phosphorylation, and Redox Status

Of the 1,038 differentially expressed genes, 17.5% (182 transcripts) were related to mitochondrial electron carrier activity, TCA cycle, oxidative phosphorylation, and redox, with 140 downregulated and 42 upregulated in the hypergravity environment (Table 2.2). Notably, genes related to all four electron transport chain (ETC) enzymes (Complexes I, II, III, & IV) and hydrogen-exporting ATP Synthase (Complex V) were highly downregulated. These robust changes may indicate impaired mitochondrial function and ATP synthesis in pupae subjected to chronic hypergravity. Anomalous ETC enzyme activity can cause electron leakage

and free radical generation in mitochondria, leading to the state of oxidative stress (Cadenas and Davies 2000). Decrease in the expression of these important genes in pupae developing in a hypergravity environment might adversely impact the process of metamorphosis.

Genes in the cytochrome P450 family (terminal oxidase enzymes in the electron transfer chain) were significantly elevated in hypergravity, suggesting a change in redox homeostasis. Increased expression of cytochrome P450 genes here could partially explain the induction of oxidative stress in hypergravity-exposed pupae due to impaired mitochondrial function and ATP synthesis. Some studies suggest that mild hypergravity exposure is beneficial to the stress response of adult flies, although no correlation with antioxidant defense, particularly SOD and catalase activity, was found (Le Bourg and Fournier 2004). Conversely, several reports have indicated that altered gravity (hypergravity/microgravity) induces oxidative stress in multiple organisms (Sajdel-Sulkowska et al. 2007). A recent study with endothelial cells found a significant decrease in expression of genes related to oxidative phosphorylation in microgravity (Versari et al. 2013). Similarly in rodents, spaceflight-induced impairments to mitochondrial function, redox status, and associated oxidative stress were reported (Stein 2002). In the same study, post-flight-induced lipid peroxidation and diminished antioxidant defense were attributed to the stress associated with reentry into the Earth's gravity, including a brief hypergravity exposure.

2.4.2 Hypergravity Affects Ion Channels / Transmembrane Transporters

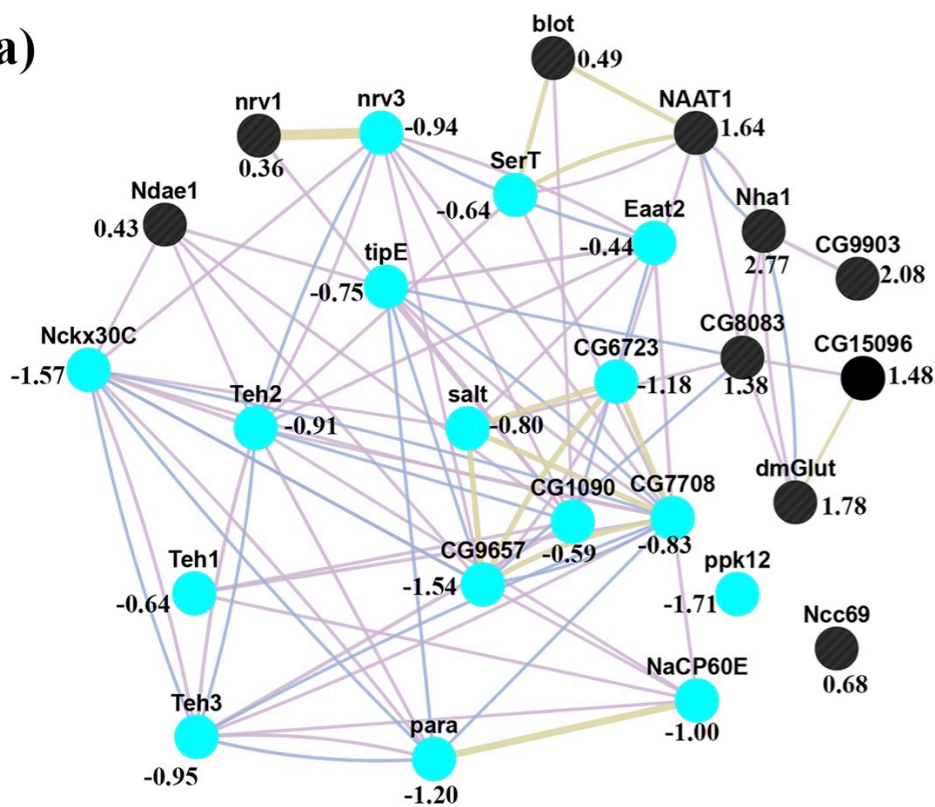
Metamorphosis in pupae involves the dramatic rearrangement of cells and changes in cell adhesion. To sense and respond to an external physical force that may affect the precision of these rearrangements, cells can activate mechanosensors. Ion channels and transmembrane transport proteins can transduce mechanical force into biochemical cell signaling pathways (Cheng Zhu, Gang Bao, and Wang 2000). In our study, GO classification revealed that expression of several ion channel and transmembrane transport protein genes were altered in response to hypergravity. Among the 158 genes (15% of the total) in this category, 94 genes were downregulated and 64 upregulated. Nearly half (73 of 158) are related to Na⁺ (26) Ca²⁺ (34) and K⁺ (13) ion channels (shown in Figure 2.4 a,b,c respectively). The majority of the transcripts related to Na⁺ and K⁺ ion channels (16 of 26 Na⁺ and 10 of 13 K⁺) were decreased in expression in hypergravity compared to control. Na⁺ and K⁺ ion channels ensure optimum action potential across the membranes of excitable cells such as neurons and myocytes (Alberts et al. 2002). A decrease in the expression of these channels in response to hypergravity could impair the action potential across the cell membrane, ultimately affecting neurotransmission and muscle contraction in pupae. We also observed that most genes related to Ca²⁺ ion binding and Ca²⁺-dependent ion channel activity were downregulated (26 of 34 Ca²⁺-related genes) whereas only 8 were upregulated (Prestin, aralar1, Trpml, Rgk1, aay, Rph, Fim, and Cib2) in response to hypergravity (Figure 2.4b).

Table 2.2 Hypergravity-induced changes in redox genes

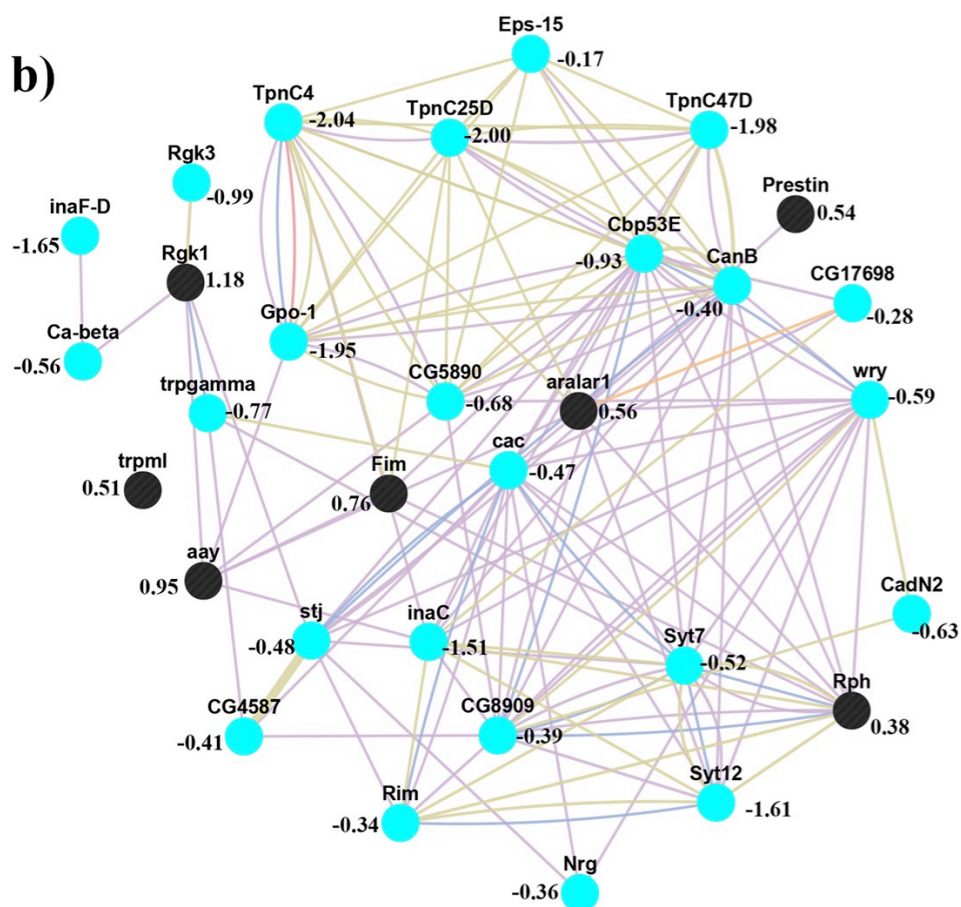
Upregulated genes			Downregulated genes		
Gene	Log2 FC*	P-Value**	Gene	Log2 FC*	P-Value**
<i>lox</i>	5.698	2.36E-43	<i>Pxd</i>	-2.956	3.90E-29
<i>CG2070</i>	4.045	2.38E-13	<i>CG30427</i>	-2.791	2.31E-12
<i>CG11878</i>	3.600	4.10E-17	<i>v(2)k05816</i>	-2.200	9.74E-06
<i>Eo</i>	3.432	1.66E-21	<i>Pdh</i>	-2.140	3.31E-05
<i>Cyp304a1</i>	2.544	4.01E-49	<i>Cyp6w1</i>	-1.955	0.000484
<i>CG4623</i>	2.502	5.77E-32	<i>CG40485</i>	-1.953	9.02E-07
<i>CG11353</i>	2.423	2.01E-22	<i>CG34172</i>	-1.953	6.32E-09
<i>Fbp2</i>	2.167	3.30E-05	<i>CG8757</i>	-1.618	0.000672
<i>CG10131</i>	1.957	6.69E-07	<i>CG8630</i>	-1.587	7.04E-05
<i>CG9914</i>	1.831	2.06E-05	<i>kdn</i>	-1.413	2.65E-05
<i>ImpL3</i>	1.662	0.000119	<i>Cyp4p1</i>	-1.371	5.97E-05
<i>Cyp6a8</i>	1.661	0.000583	<i>l(2)01289</i>	-1.365	8.60E-05
<i>CG12268</i>	1.637	0.000711	<i>CG5955</i>	-1.332	1.00E-06
<i>mex1</i>	1.595	3.24E-05	<i>CG9150</i>	-1.273	0.002439
<i>CG9518</i>	1.590	0.001960	<i>Cyp6a9</i>	-1.206	0.004563
<i>CG9747</i>	1.582	0.001417	<i>CG4769</i>	-1.201	5.24E-06
<i>GILT2</i>	1.566	0.001001	<i>Cyp4g15</i>	-1.163	2.88E-05
<i>Spat</i>	1.562	0.000268	<i>l(1)G0156</i>	-1.055	5.95E-05
<i>CG10211</i>	1.552	0.001267	<i>CG5028</i>	-1.041	6.08E-05
<i>CG9512</i>	1.538	0.000275	<i>Cyp311a1</i>	-1.027	-4.10357

Top 20 upregulated and downregulated genes related to oxidation-reduction, electron transport chain, and TCA cycle in pupae that developed in hypergravity compared to 1 g. *Upregulated and downregulated genes are indicated as positive and negative fold change respectively. **P-value adjusted for FDR = 0.05.

a)



b)



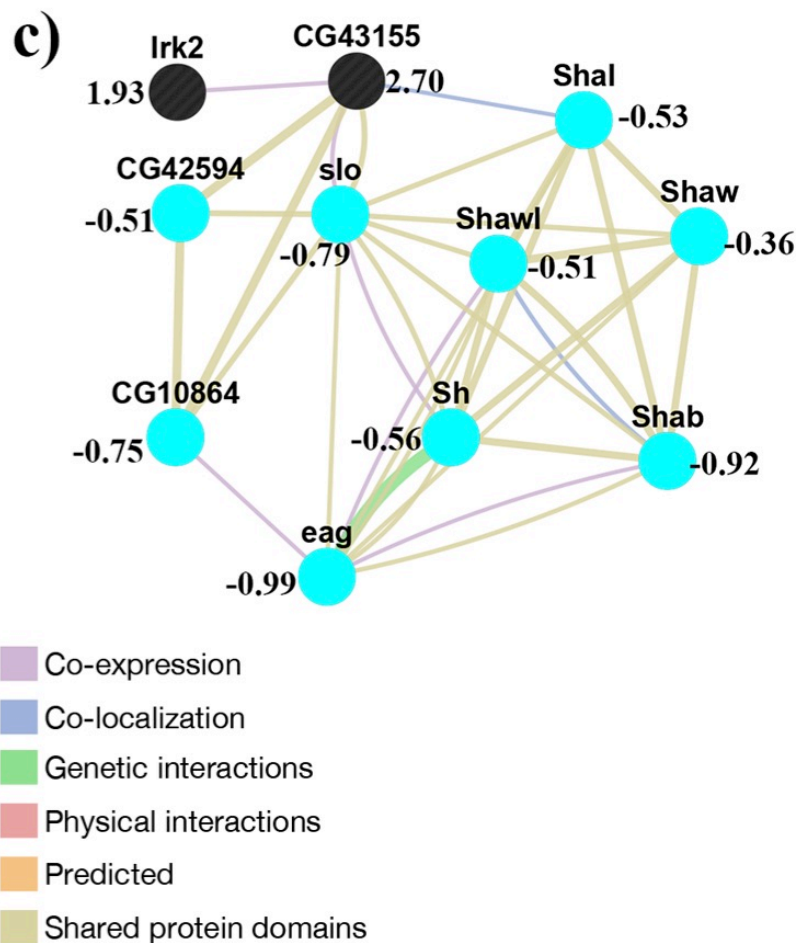


Figure 2.4 Hypergravity-induced changes in transport genes

Nearly 15% of the differentially expressed genes were related to ion and transmembrane transport. Transport genes have been further subdivided into sodium-related (a, previous page), calcium-related (b, previous page), and potassium-related (c, this page) transport genes. Blue – Downregulation; Black – Upregulation.

Mechanical loading such as hypergravity can activate mechanotransduction pathways. Several reports suggest roles of mechanoreceptors in sensing gravitational acceleration (Burighel, Caicci, and Manni 2011). In mammals, mechanical stimuli play a vital role in bone remodeling, especially bone cell differentiation, proliferation, and maturation, which are extremely sensitive to mechanical cues (Kang et al. 2011). Similarly, mechanotransduction pathways regulate root gravity-sensing cells and calcium homeostasis in plants (Kolesnikov et al. 2015). However, it is not yet clear exactly how altered gravity manipulates transport proteins and ion channel activity. One possibility is that interaction between integral membrane proteins or the plasma membrane itself changes in gravitational acceleration. Goldermann and Hanke (2001) supported this hypothesis by demonstrating changes in native ion channels of the *E. coli* outer membrane in response to altered gravity (Goldermann and Hanke 2001). Furthermore, they confirmed that changes in acceleration can alter the structural and

functional properties of alamethicin, a synthetic polypeptide that forms voltage-dependent ion pores in artificial and biological membranes (Goldermann and Hanke 2001). These simplified membrane studies indicate detection of changes in gravitational acceleration might be an intrinsic property of biological membranes. Thus, changes in mRNA transcripts related to ion channels and transport could be part of the mechanotransduction-mediated manipulation causing metamorphosing cells in pupae to actively adapt to the change in gravitational acceleration.

2.4.3 Hypergravity Elicits Immune Response in Pupae

It is widely recognized that altered gravitational force has a profound influence on immune function. Spaceflight and ground-based studies validate this in different organisms (Sonnenfeld and Shearer 2002; Crucian et al. 2012; Guéguinou et al. 2012; Marcu et al. 2011). In the present study, one major category of differentially expressed genes related to immune pathways. 5% of differentially expressed genes were immune-related genes that were highly upregulated in response to the chronic hypergravity environment (Figure 2.5). The expression of several ecdysone-induced genes with dual roles in both metamorphosis and defense response against bacterium were significantly increased, suggesting a role of ecdysone hormone in regulating innate humoral immune response. Juvenile hormone (JH) and 20-hydroxy-ecdysone (20E) play critical roles in metamorphosis, and evidence suggests that both JH and 20E modulate innate immune response as well (Flatt et al. 2008). Related to this, when we increased hypergravity to 8 g, we discovered not only accelerated pupal development but also a significant reduction of adult eclosion. This phenotype may partially be attributed to robust changes in the expression of ecdysone-induced genes in response to hypergravity. Thus, it is likely that hypergravity not only alters hormonal regulation, affecting metamorphosis in pupae, but also modulates innate immune response. We also found increased RNA levels in specialized immune-related genes including Mtk, Drs, Drsl1, Drsl6, PGRP-SA, PGRP-SB1, PGRP-SB2, spz3, IM2, IM3, ird1, Pirk, and AttB in response to hypergravity, suggesting a direct impact on innate immune function. In the absence of microbial infection, increased expression of antimicrobial peptides indicates an immune boost in response to hypergravity. Additional studies support our finding that hypergravity boosts immune function. Hypergravity-exposed flies (wild type and immune-compromised) exhibited increased survival against fungal infection (Taylor et al. 2014). In addition, in the absence of microbial infection, Lee et al. have shown that mechanical loading and stress activate immune responses in rat and human cell lines (Lee et al. 2012). In contrast, microgravity exerts the opposite response by compromising immune function. For instance, adult flies returned from space showed increased vulnerability to microbial infection (Marcu et al. 2011; Taylor et al. 2014). Similarly, previous studies from our group also show altered expression of antimicrobial peptides in response to space flight in both larvae and adults (Marcu et al. 2011). Hypergravity may have an advantage in terms of host immune defense response to combat against microbial infection.

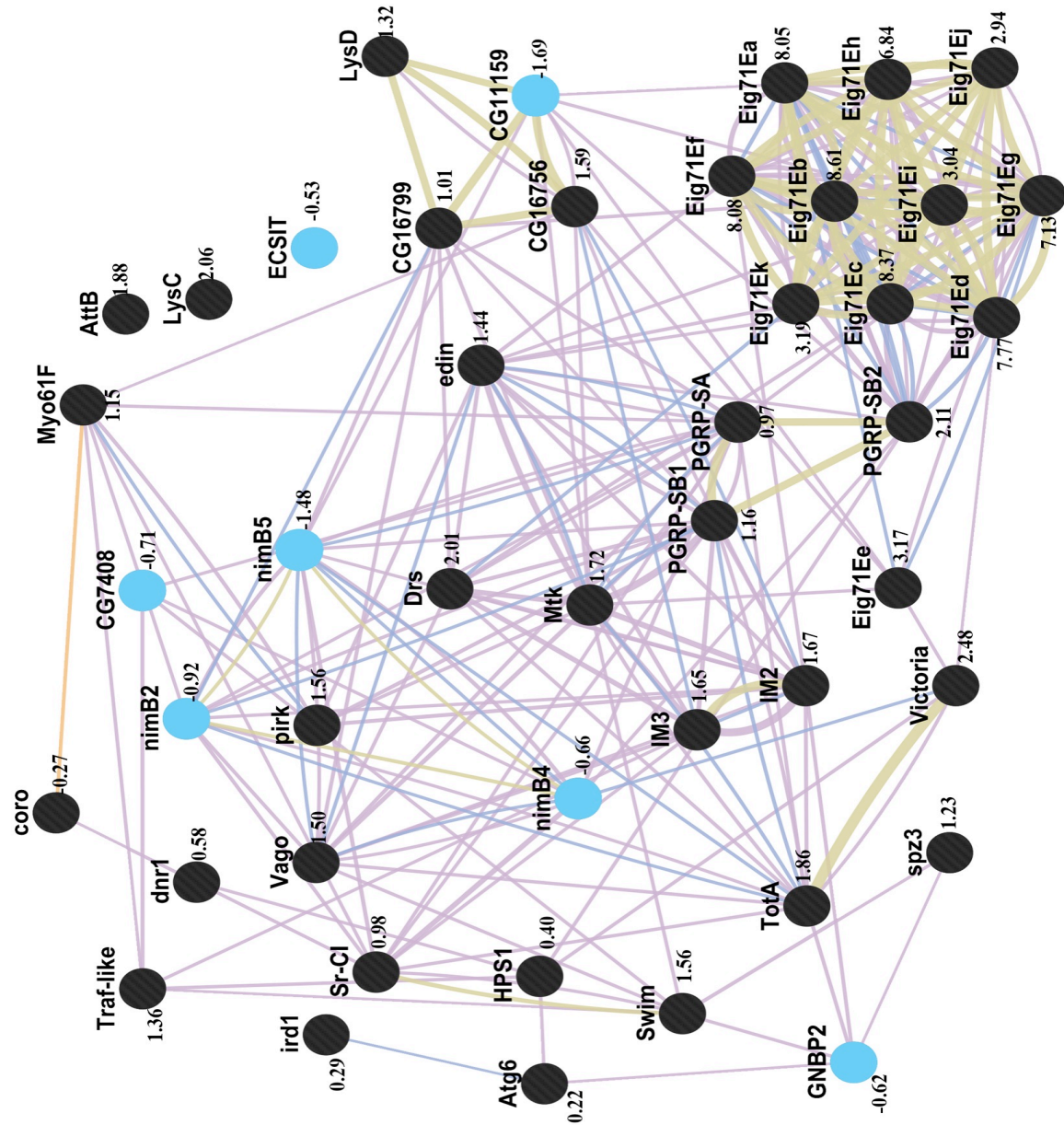


Figure 2.5 Hypergravity affects defense genes

Among the 1,038 differentially expressed genes in pupae developed in hypergravity, nearly 5% are host-defense related genes. These host defense genes relate to microbes, Gram-positive and Gram-negative bacteria, fungus, and virus. Blue – Downregulation; Black – Upregulation. For gene network keys see Figure 2.4.

metabolic processes and 43 chitin-based cuticle development genes were altered in response to hypergravity (Figure 2.6). Chitin plays a major role in structural integrity by stabilizing the layered organization of the cuticle (Moussian et al. 2005). Changes in the expression of these genes suggest altered chitin binding and metabolism. This data corroborates well with previous findings, by Herranz et al., that genes related to insect cuticle proteins were significantly altered during metamorphosis, with most being downregulated (Herranz et al. 2013). In insects, cuticle serves a function similar to that of the skeletomuscular system in mammals. However, unlike the mammalian model, not many studies have been carried out in insects to understand the effect of altered gravity on cuticle development and chitin biosynthesis. One study, carried out on the space shuttle Challenger in 1985, found a pronounced decline in larvae hatched from the embryonic cuticle, as well as alterations in larval head, anterior, and thoracic cuticle segments (Vernos et al. 2002). Our research from Space Shuttle Discovery (STS- 121, 2006) data revealed that defects in the pupal cuticle correspond with decreased eclosion rate. Many of the flies attempting eclosion rather died half-eclosed from the pupal case, suggesting that a structural change in the pupal case during spaceflight prevented the flies from properly eclosing (Hosamani, unpublished observations). The significant defects in the embryonic, larval, and pupal cuticle echoes the skeletomuscular system of mammals: like skeletomuscular tissues, insect cuticle is highly vulnerable to microgravity/spaceflight effects.

In the present study, overrepresentation of cuticular genes due to hypergravity lends support to the phenotypes observed in the spaceflight experiments and suggests candidate genes affecting cuticle development in the pupae exposed to microgravity.

2.4.5 Proteolysis Altered in Hypergravity

During metamorphosis, active restructuring of organelles occurs in the early pupa. A perfect balance between protein degradation and synthesis must occur during this period (Udvardy 1993). Endopeptidases catalyze protein degradation, freeing amino acids that in turn can be recycled for protein biosynthesis. Changes in proteolytic gene expression (65 upregulated and 24 downregulated) in hypergravity may indicate an imbalance in protein biosynthesis.

2.5 Additional Transcriptional Response in Hypergravity

59 genes encoding mitochondrial ribosomal proteins were significantly downregulated. 37 large subunit (mRpL) and 22 small subunit (mRpS) ribosomal protein encoding genes were decreased in their expression in pupae developed at 3 g. Mitoribosomes are located in the mitochondrial matrix and synthesize proteins involved in oxidative phosphorylation (Lightowers, Rozanska, and Chrzanowska-Lightowers 2014). The expression of genes coding for mitochondrial electron carrier activity, oxidative phosphorylation, and redox homeostasis were the largest group of

genes altered in this study. Decrease in the expression of these genes could cause a cell-autonomous defect in protein biosynthesis in hypergravity.

Nine genes with isomerase activity related to cuticle pigmentation and melanin biosynthetic processes were altered in the hypergravity environment in pupae. Genes related to carbohydrate, lipid metabolic processes, protein glycosylation, glycolipid biosynthetic processes, catecholamine metabolism, acetyl-CoA biosynthetic processes, xenobiotic metabolism, ketone body catabolic processes and oligosaccharide biosynthetic processes were also altered in pupae developed in the hypergravity environment.

2.6 Hypergravity Versus Pro-Oxidant-Induced Transcriptional Response

It has been reported that hypergravity exposure is associated with oxidative stress (Versari et al. 2013). We sought to discern which gene expression changes in our study were specific to hypergravity and which overlapped with other oxidative stress responses triggered by common pro-oxidants. We compared our data with Girardot et al., who studied transcriptional responses to common pro-oxidants such as paraquat and hydrogen peroxide (Girardot, Monnier, and Tricoire 2004). While the Girardot et al. study used adult flies and our study deals with early pupae, oxidative stress response is conserved at a molecular level across the developmental stages and across species (Davies et al. 2014; Feder and Hofmann 1999). With this caveat, we compared our data to Girardot's results. Genes that were affected by hypergravity and encode mitochondrial electron transport chain, redox, and oxidative phosphorylation proteins significantly overlapped with those induced by paraquat and hydrogen peroxide exposure. Chronic hypergravity might be inducing reactive oxygen species (ROS)-mediated oxidative stress in pupae similarly to pro-oxidants in adults. In contrast, genes encoding ion channel and transport proteins, cuticle development, and immune response overlapped the least with previously published data. This preliminary comparison of transcriptomic data suggests hypergravity may induce oxidative stress comparable to many of these pro-oxidants, but that also it has specific effects on mechanotransduction, cuticle development, and immune response.

To assure that these specific effects are unique to hypergravity, comparison of our data to environmental stress-induced transcriptional responses reported by Brown et al. 2014 show that, while hypergravity response is slightly correlated to other stressors (Spearman between 0.27 for heat shock and -0.033 for 5mM paraquat), the main responses elicited are particular to hypergravity (Figure A.1). 782 of our genes were altered under stress conditions in Brown et al., including broad upregulation of cytochrome P450s across treatments (Supplemental Table S6 of published edition). Correlation between our samples and the study's "homogenous response to environmental stressors" could be affected by the latter's use of adult flies verses our use of pupae.

3. Conclusions

We performed RNA-Seq on pupae in hypergravity to determine the most important factors in transcriptional response to abnormal gravity. We found 1,513 genes differentially expressed in response to hypergravity, with 1,038 of the genes falling into 13 functional categories. Using qRT-PCR, we were unable to find a linear relationship between transcript levels and magnitude of g force, possibly because higher g-levels caused non-specific stress and decreased viability. However, we did find that increased gravitational force increases the rate of development. As the expression profile for hypergravity deviates from other known stress expression profiles, we conclude that this rate increase is a gravity-specific response. Further studies will inform this observation.

In planning this study, we expected to see changes in specific developmental pathways in *D. melanogaster* pupae since we had previously observed an increase in pupal lethality after spaceflight (Bhattacharya, unpublished data). We found a wide array of biological processes altered in response to hypergravity exposure. A possible explanation for this diverse transcriptomic response lies in the evolutionary adaptation of an organism like *Drosophila melanogaster* to Earth's gravity. Organisms are constantly sensing and adapting to Earth's different environmental cues. Specific genes accordingly have evolved in concordance with a constant gravitational force of 1 g. Orienting themselves with respect to Earth's gravity is an essential act of most multicellular organisms including plants, arthropods, and chordates. It is to be expected therefore that changes in gravity elicit responses in multiple pathways of an organism. As a result, the transcriptome in its entirety may respond quite diversely to cope with this change in gravitational force of three times its normal level. Thus, we see hypergravity affecting a broad range of biological processes in *Drosophila melanogaster* pupae.

4. Materials and Methods

4.1 *Drosophila* Culture and Maintenance

Canton-S wild type flies were cultivated at 25 °C and 50% relative humidity on standard fly food (torula yeast 25.6 g; dextrose 103.2 g; cornmeal 48.8 g; agar as a solidifying agent 7.44 g; and tegosept as an antifungal 18 mL/L H₂O) with 12 h light–dark cycles. To generate synchronized 2 to 3 day old F1 flies, adult flies were allowed to lay for 48 h in a bottle, then parent flies were discarded, and the bottles with eggs were allowed to grow into adult flies at 25 °C. From the F1 generation, 2 to 3 day old flies were used for the chronic hypergravity studies using standard laboratory fly vials.

4.2 Hypergravity Exposure

We define our hypergravity condition as 3 times the Earth's gravitational acceleration (3 g). Hypergravity was mimicked using the standard centrifugation method (using a modified BECKMAN GS-6R centrifuge designed to attain low g levels – 97 RPM is equal to 3 g). For chronic hypergravity exposure, 20 female and 20 male synchronized 2 to 3 day old flies were allowed to lay eggs while being maintained in the 3 g environment. Centrifugation was briefly stopped to discard the parent flies after 24 h of egg laying, and then immediately continued until the eggs developed into P6-staged pupae. P6-stage of pupae typically lasts between 25 and 43 h after pupation initiation (white puparium stage). We ascertained this stage by isolating pupae at 40 h after pupation, which is well within the P6 stage. The development of green malpighian tubules is the hallmark of this stage, and we confirmed this by dissecting representative pupae. Similarly, 1 g controls were maintained in the dark, with the same angle of vial orientation so as to mimic the hypergravity-exposed flies. It is not possible to sex pupae at stage P6, as the differentiating male sex combs do not darken and become visible until stage P12 (Tyler 2000). We considered sexing at the larval stage by looking for the enlarged male gonad and then separating the sexes into separate vials to develop, but we were concerned that the time taken to do this would interrupt the prolonged 3 g-exposure environment we were creating. Therefore we used downstream statistical methods to confirm that our gene expression results correlated with hypergravity treatment and not to the sex of the organism, as is described in Section 2.2.

4.3 Total RNA Isolation and RNA-Seq Library Preparation

Post hypergravity exposure, P6-staged early pupae were collected. Using Qiagen RNeasy mini kit, total RNA was extracted from four pooled pupae for both conditions (3 g and 1 g), and in addition four such independent biological replicates were processed to ensure statistical reliability. Total RNA for these 8 samples (4 from each condition) were treated with DNase (turboDNase) and prepared for sequencing (Illumina mRNA TruSeq vol. 2 kit) as per the manufacturer's protocol. Size distribution and concentration of DNA fragments were measured using chip electrophoresis (Agilent Bioanalyzer) and fluorometer (Qubit 2.0). Indexed libraries were pooled (8 samples total) and sequenced on one flowcell lane (Illumina HiSeq 2000) to obtain 100-base paired-end reads.

4.4 Transcriptome Analysis

Raw sequenced reads were filtered to remove reads not passing Illumina's Casava quality score, and then trimmed (Trimmomatic) to remove adapters and low-quality reads. Trimmed reads were aligned to the *Drosophila melanogaster* reference transcriptome build r6.02 (Bowtie 2, default settings) Table 2.3 details reads for each sample. Aligned reads were subsequently quantified (eXpress). The transcript TPMs

were merged to gene level and scaled up ($\times 100$) for downstream analysis. Differential expression analysis was performed (DE-Seq2) to measure fold change and significance between 1 g and 3 g conditions. Genes with q-values ≤ 0.05 were considered significantly differentially expressed. Pathway analysis was performed to determine significant affected processes (PANTHER). Using the Gene Ontology hierarchical structure, where specialized ‘child’ terms are nested within broader ‘parent’ terms, we grouped terms that had the same molecular functions and biological processes. PANTHER statistical overrepresentation tests for biological processes and molecular functions were performed with the control samples' genes as reference. PANTHER statistical enrichment tests were performed for biological processes and molecular functions using default settings. All statistical tests were corrected for multiple testing using the Bonferroni method. Results of these tests are detailed in Appendix Tables A.1-A.4. Gene network figures were generated using GeneMANIA (Warde-Farley et al. 2010). These networks provide information on genetic interactions, physical interactions, co-expression, co-localization, and shared protein domains between genes. Software versions and settings listed in Table A.6.

Table 2.3 Read details for samples

Sample	Index	Adapter	Raw Reads (Paired)	Filtered Reads	Trimmed Reads	Percent Aligned
1GR1	16	CCGTCC	27199285	24457692	24424375	92
1GR2	4	TGACCA	33305499	30090146	30050069	93
1GR3	5	ACAGTG	26332294	23850461	23818861	92
1GR4	6	GCCAAT	33587499	30248380	30206827	91
3GR1	7	CAGATC	29962438	26991097	26955295	93
3GR2	12	CTTGTA	30136115	27189981	27152866	92
3GR3	2	CGATGT	26893693	24332694	24299836	92
3GR4	18	GTCCGC	30383814	27347784	27310536	93

4.5 qRT-PCR

Based on GO analysis, we selected 18 genes representing various pathways for validation by qRT-PCR. Relative gene expression levels were quantified (7500 Real Time PCR instrument from Applied Biosystems, and SYBR green PCR kit, QuantiTect, Qiagen). Primers were designed (Primer-Blast, NCBI). Real-time cycler conditions were: 95 °C for 15 min for initial activation step, and 40 cycles of 94 °C for 15 s, 50 °C for 30 s, and 72 °C for 30 s for denaturation, annealing, and extension respectively. Data was normalized to the reference gene alpha tubulin 84B. The comparative cycle threshold (CT) method was used for data analysis. Primers used for target genes listed in Table A.7. Results for qRT-PCR tests on samples exposed to 3 g, 5 g and 8 g are provided in Table A.5. Exposure to 3 g, 5 g and 8 g hypergravity levels were conducted as described previously in Section 4.2 and corresponded to 97

rpm, 161 rpm, and 258 rpm settings respectively on the modified BECKMAN GS-6R centrifuge.

4.6 Availability of Data and Material

The datasets supporting the conclusions of this article have been deposited in NCBI's Gene Expression Omnibus (Edgar, Domrachev, and Lash 2002) and are accessible through GEO Series accession number GSE80323 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80323>). Other supporting material is available within the published article and in the Appendix. Author contributions are detailed in the published article. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303.

3 Fast Fusion Detection, Assembly, and Quantification from RNA Sequencing Data

I. Introduction

Accelerated acquisition of heterogeneous, somatic mutations across cells within a tumor or pre-tumorous tissue is a widespread occurrence in most cancer types (Hanahan and Weinberg 2000). Resultant mutations are sufficient to induce tumorigenesis and drive tumor progression and metastasis through activation of proto-oncogenes and deactivation of tumor suppressor genes. The accumulation of mutations can be induced through the malfunction of several biological pathways, involving several corresponding mutation classes (Stratton, Campbell, and Futreal 2009). One common and particularly deleterious class of mutation is chromosomal rearrangement. Chromosomal rearrangements typically occur when two or more double-strand breaks are incorrectly repaired through rearrangement, deletion, or duplication. Large-scale changes to gene expression from such rearrangements are associated with a number of disorders (Mertens et al. 2015). When the rearrangement breakpoints are at gene regulatory and/or coding regions, the disruption can result in a deregulated or a *chimeric* fusion gene (Figure 3.1). Gene fusions are associated with many types of cancer and play a major role in tumorigenesis.

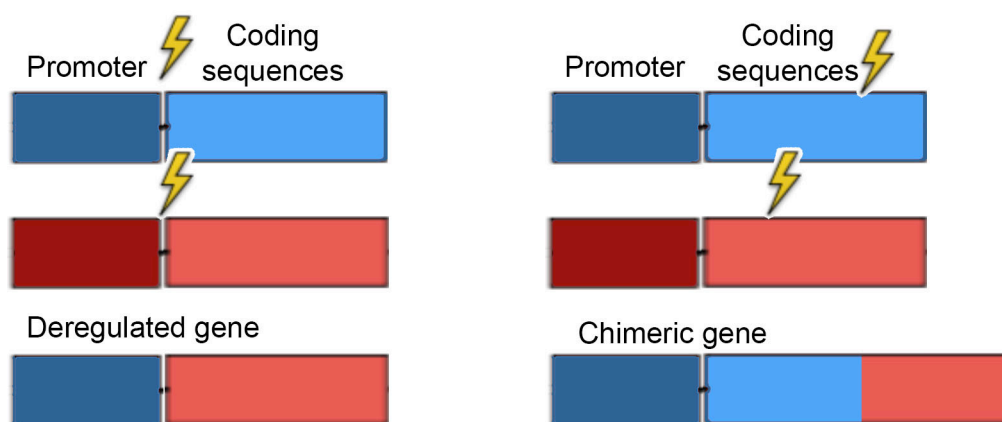


Figure 3.1 Fusion genes resulting from chromosomal rearrangements

If the chromosomal rearrangement is made at coding and/or regulatory regions, a deregulated or chimeric gene can occur. Lightning bolt indicates chromosomal breakpoint.

Specific gene fusions often correspond to specific tumor types. Recurrent gene fusions occur at fragile genomic loci during tumorigenesis in certain leukemias, melanomas, and carcinomas (Yunis and Soreng 1984). In chronic myeloid leukemia (CML), a fusion between breakpoint cluster region (BCR) and Abelson murine leukemia (ABL) virus genes leads to the recurrent BCR-ABL fusion which is present in a large percentage of CML patients (Rowley 1973). This fusion is reported to

transform healthy cells into cells with tumor characteristics, and has been successfully targeted by Imatinib, one of the world's first successful targeted therapies in terms of extending patient overall survival time (Chin, Andersen, and Futreal 2011). Several similar targetable gene fusions have been discovered. Because of this fusion-tumor specificity, gene fusions are useful for cancer diagnostics and therapeutic drug targeting (Mertens et al. 2015). One goal of major consortium efforts such as The Cancer Genome Atlas (TCGA; Weinstein et al. 2013) and The Cancer Cell Line Encyclopedia (CCLE; Barretina et al. 2012) is to detect recurrent fusion genes from different cancer types based on sequencing information. Recently, for example, recurrent receptor tyrosine kinase fusions were detected in gliomas (Cancer Genome Atlas Research Network et al. 2015). These consortia hope to better characterize cancer-associated fusions and to aid in developing additional targeted cancer therapies.

Currently, most fusion detection methods rely upon RNA sequencing (RNA-Seq) (Ozsolak and Milos 2011) because it is cost-effective, covers all expressed regions of the genome, and is highly interpretable. Several computational tools to detect fusions from RNA-Seq reads exist. Fusion detection tools typically operate via three steps: First, detect “spliced” reads where the first part of the read maps to one gene and the second part maps to another gene or detect “spanning” paired-end reads where one entire read maps to one gene and the paired-read maps to another gene. Second, aggregate reads into candidate fusion junctions through realignment-based grouping. And last, filter candidate fusions based upon a number of heuristics (Kumar, Razzaq, et al. 2016).

In practice, identification of gene fusions via RNA-Seq is difficult and error-prone. While over 20,000 gene fusions have been identified in the three major neoplasia subtypes, fewer than a 1,000 of them have been confirmed as recurrent (Mertens et al. 2015). Factors contributing to the inflated number of positive fusions include incorrect read mapping, transplicing, and template switching. Incorrect mapping is often a result of fusion detection tools attempting to map fusions from transcripts arising from repetitive regions of the genome such as germline segmental duplications. Transplicing can induce read-through events, where transcription continues across genomically adjacent genes to be transcribed into one RNA transcript. Template-switching events also contribute to the count of false positive gene fusions as the RNA polymerase jumps from one DNA template to another during transcription (H. Li et al. 2009). These produce low but detectable baseline levels of fusion genes in wild-type as well as cancerous cells, but are usually not of interest in cancer sequencing efforts as they are not causally involved in tumorigenesis, cancer progression, or metastasis (Gingeras 2009).

To filter out many of the false positives, fusion detection tools employ a set of heuristics based upon the contributing biological factors just described, as well as a number of ad hoc methods. Existing fusion detection tools each contain some mixture of the following disadvantages: many hours to run, computationally expensive, poor sensitivity/specificity, nontrivial to install, questionable heuristics, no breakpoint resolution, no fusion quantification.

With heuristics, there is a trade off between type 1 and type 2 errors. For instance, by ignoring fusion genes originating from repetitive genomic regions, a tool may filter real fusions that are important to cancer characterization (Kumar, Razzaq, et al. 2016). It is an ongoing issue in fusion genomics how best to combine heuristics in a way generalizable to the majority of fusion gene discovery use cases. This is due in part to the wide range of genomic instability that tumors from different tissues have exhibited. As a result, machine-learning based classifiers tuned on representative datasets behave poorly when introduced to new tumor types (Beaumeunier et al. 2015). Additionally, fusion identification that relies upon first-generation alignment algorithms is slow and computationally intensive because, in most methods, every read must be split in a number of ways and positions and matched against all possible regions of the genome through alignment during chromosomal rearrangement-finding steps (Kumar, Razzaq, et al. 2016).

Here, we introduce an improved method for gene fusion discovery utilizing the publicly-available software kallisto (Bray et al. 2016), based on pseudoalignment, along with a downstream processing tool that (1) filters false positives using biologically-relevant heuristics, (2) assembles transcripts at breakpoint resolution, and (3) quantitates fusion abundance. This method is computationally and temporally efficient, allowing for laptop-based analysis, clinical use, and investigation of large cancer gene expression databases to probe for tumor-specific and cancer-type agnostic recurrent fusions.

2. Methods

Fusion junctions are detected using a two-stage method. The first stage is implemented in kallisto and detects individual reads or read pairs, whose parts individually pseudoalign to the reference transcriptome but not when combined together. Pizzly takes the output of the potential fusion junctions found by kallisto and does a more detailed analysis of the reads by accurately aligning the reads across the junction. Additionally, pizzly is annotation aware, i.e. it uses information about the genomic coordinates and gene identities of each isoform when making a fusion junction call.

2.1 kallisto Fusion Stage

The kallisto program was modified to search for reads that could support fusion junctions. Kallisto uses a k-mer based index for the reference transcriptome when computing a pseudoalignment for reads. For each k-mer the index records the set of transcripts containing this k-mer, which is its equivalence class (EC). For normal reads arising from a single transcript, all matching k-mers are supported by at least that transcript and the intersection of the ECs for this read will be non-empty and contain the true transcript. For reads or read pairs that span a fusion junction, the

ECs from each side will have an empty intersection and thus be discarded from further consideration by kallisto. In the modified version, we specifically look for read pairs discarded because the intersection of ECs is empty, yet some k-mers matched. Such reads are kept as part of the fusion output when either one of the following holds: (1) each read has a non-empty EC intersection separately, but combined the intersection is empty (this is consistent with Figure 3.2 “paired fusion reads,” where the reads come from opposite sides of the fusion junction) or (2) one of the reads can be split into two parts such that the first part of the read has a non-empty EC intersection and the remainder of the read, along with the other read from the pair, has a non-empty EC intersection (this is consistent with Figure 3.2 “split fusion reads,” where one of the reads spans the fusion junction). When a potential split of the reads has been identified, we check all matching k-mers and require further that the union of ECs on either side is empty. This is to lower the number of false positives due to reads from unannotated transcripts that look like fusions between related transcripts. All read pairs matching these criteria are saved along with supporting information about the matching transcripts.

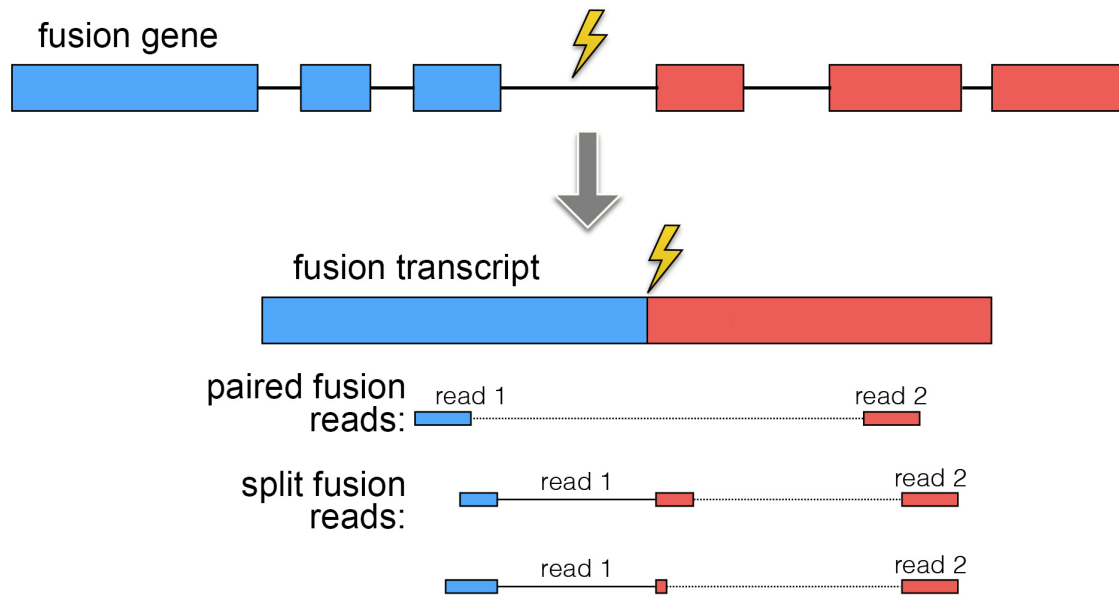


Figure 3.2 Reads meeting specific criteria are passed to pizzly

A fusion gene, resulting from chromosomal rearrangement, is transcribed into a fusion transcript. Standard kallisto does not pseudoalign reads spanning the fusion junction to the transcriptome. In the fusion-modified kallisto, reads meeting specific criteria are passed to pizzly as fusion candidates.

2.2 *pizzly* Fusion Stage

Unlike kallisto, which is largely annotation agnostic, pizzly performs a more detailed analysis of the reads to determine fusion breakpoints by taking the annotation of the transcriptome into account. Pizzly accepts the annotation in GTF format, where it

obtains the exact genomic locations of exons and whether or not the transcript is annotated as protein coding.

The input to pizzly is the set of read pairs that kallisto flagged as potentially spanning fusion junctions. In the first step of pizzly, each read is evaluated independently and reads that are classified as false positives are rejected. Pizzly uses several criteria to reduce the number of false positives. First, reads that map to multiple genomic locations are discarded. Next, since kallisto is unaware of which transcripts belong to which genes, its output contains many false positives where the potential fusions are between two distinct transcripts from the same gene. These are typically from transcript isoforms that were not present in the reference transcriptome. These are discarded. The output of kallisto also contains many false positives where read pairs originate from a single transcript with mismatches (SNP, etc.) to the reference, but kallisto assigned it to two distinct but similar genes. Rather than relying on annotation of gene families to filter these cases, we filter them using approximate sequence alignments. We consider the matching k-mers of one end and look for approximate matches to the transcripts of the other end (Figure 3.3). Rather than aligning the k-mer matches to the entire transcriptome with mismatches, we only consider the potential candidates kallisto identified for the other end. If any such approximate match is found for either one of the ends, the read pair is discarded.

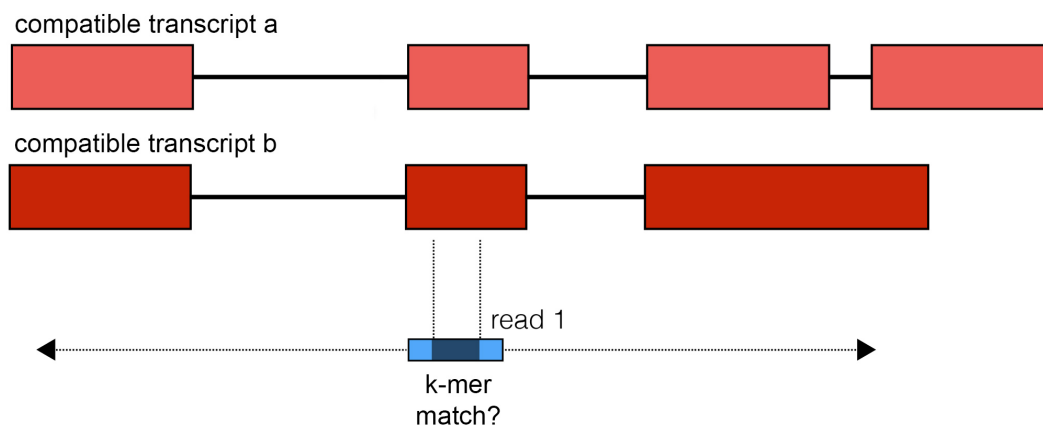


Figure 3.3 A check for alignment to partner transcript ensures correct fusion

The k-mers from the first part of the candidate fusion, the blue “read 1”, are aligned with mismatch allowance to all compatible transcripts from the second part of the candidate fusion. If a match is found, this false positive is discarded. The same is repeated for the other end of the candidate fusion.

Third, we require the entire read to align with mismatches to the transcripts identified by kallisto. If each read aligns to representative transcript sequences, we label the read pair as “paired” (Figure 3.2). In cases where either read cannot be aligned to the potential transcript, we attempt to compute a split-read alignment (Figure 3.4). In order to limit the potential for false positive splits, we restrict the search space to only include split sequences that fit within a specified insert length (Figure 3.5). Reads that can be split-aligned are labeled as “split.”

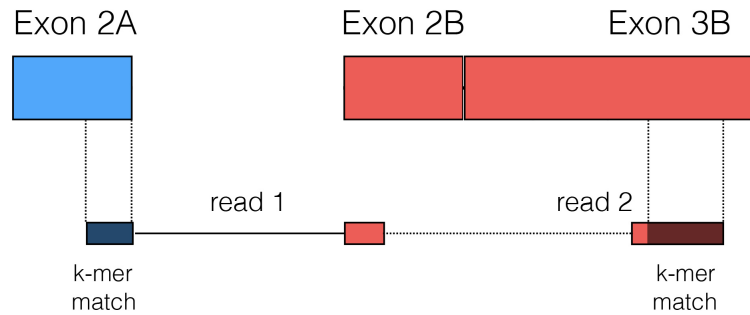


Figure 3.4 A read can be split-aligned between the two fusion transcripts
Split-read alignments provide breakpoint sequence resolution.

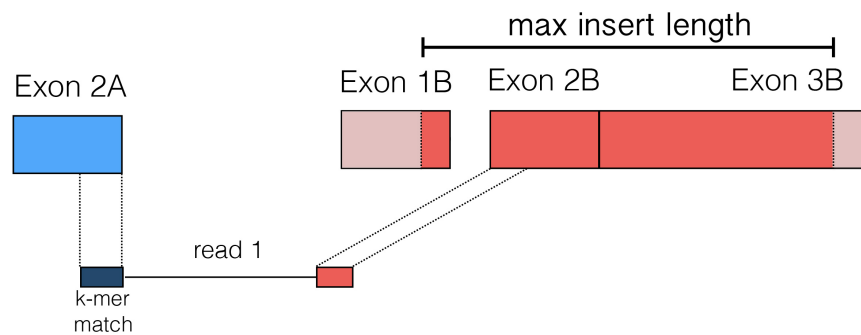


Figure 3.5 Split alignment search
The search space for split alignments is limited to the matching part of the compatible transcript up to the fragment insert length.

Finally, when all the reads have been filtered, we aggregate the information on a gene-to-gene fusion level, reporting the number of paired and split reads supporting the fusion junction. Additionally, we report each potential transcript junction, the number of read pairs supporting the transcript-level fusion as well as sequence of the fused transcripts.

3. Performance Assessment

We applied our method to multiple datasets to assess the precision/recall, computational cost, time, and ease of use of pizzly for fusion detection. To benchmark our performance, we compared our results to leading available fusion software packages, using the same datasets to evaluate the performance of each program.

3.1 Datasets

We assessed performance on four types of data to represent a range of challenges in gene fusion detection.

Positive Control

Two positive synthetic datasets of different read lengths were used, both consisting of reads from the same 50 fusion transcript sequences:

- a) A positive dataset containing 57,209 75nt paired-end synthetic reads corresponding to 50 fusion transcripts, with 4,300 reads directly covering the fusion junctions. This dataset was created by the FusionMap developers (Ge et al. 2011) and used in a previous fusion tool evaluation (Kumar, Vo, et al. 2016).
- b) A positive dataset containing 200,000 100nt paired-end synthetic reads corresponding to the same 50 fusion transcripts, simulated using the RSEM read simulator (B. Li and Dewey 2011).

Negative Control

A negative dataset containing 30 million 100nt paired-end synthetic reads simulated using RSEM read simulator using as background expression sample SRR066679 originating from H1 human embryonic stem cells (Raney et al. 2011).

Real Data with Validation

Fusion detection may perform well on simulated datasets and yet behave differently when presented with real data generated via RNA-Seq. We used real data to assess performance under standard experimental conditions.

An mRNA-Seq dataset, SRR1659964, consisting of 93,867,189 100nt paired-end reads and containing sequence from nine synthetic fusion transcript RNA constructs. This dataset was generated by Tembe et al. 2014 by pooling the nine fusion constructs at equimolar concentration and adding $-6.17 \log_{10} \text{pMol}$ to a 1 μg aliquot of total RNA, preparing an Illumina TruSeq stranded mRNA library, and sequencing on an Illumina HiSeq 2500.

3.2 Tools for Benchmarking

There are several fusion detection software packages available. We chose to focus our comparison on software that met the following criteria: (1) used in the field (based upon citation number or newness), (2) high performance as evaluated in earlier benchmarking papers (Kumar, Vo, et al. 2016; Liu et al. 2016), and (3) ability to run the software, which is not always a trivial task. The tools meeting these criteria are described in Table 3.1 below.

Table 3.1 Fusion detection tools compared to pizzly for benchmarking

tool	version	parameters	reference used
MapSplice	2.2.1	default	ensemble hg38
EricScript	0.0.5	default	E.S. formatted hg38
FusionCatcher	0.99.6a	default	F.C. formatted hg38
JAFFA	1.07	default: hybrid on 75bp, direct on 100	UCSC hg38, provided by JAFFA
PRADA	1.2	suggested -junL 80% of read length	PRADA formatted hg37
STAR-Fusion	0.8.0	default	S.F. formatted GRCh38v23

3.3 Results

Tools were run using four cores where possible. MapSplice was run on an Amazon Elastic Compute Cloud instance Ubuntu machine type c4.4xl (16cpu, 30GiB memory). PRADA was run with the suggested length of constructed junctions (-junL) of 80% of read length, which for the 75nt dataset was 60 and for the 100nt dataset was 80. JAFFA has three run modes: Assembly, Hybrid, and Direct, which are suggested for different read lengths. For short reads JAFFA assembles the reads to search for fusions in assembled contigs. For longer reads, JAFFA suggests no assembly. We followed the suggested modes for read lengths of 75nt and 100nt.

The results of the tests are summarized in Tables 3.2-3.5. PRADA failed to run on all 100nt datasets, finding fusions only for the 75bp dataset. FusionCatcher, on first examination, appeared to be working well. However, we discovered that the developers hardcoded known true and false positive fusions into the program, which allows them to more accurately predict fusions than would be possible through their program alone. While hardcoding known fusions may be helpful in a clinical setting, in exploratory analysis, such as is performed in research, this is a limiting restriction and reduces the utility of the program. We advise readers to take this into account in the benchmarking results. EricScript predicted true fusions decently, but had difficulty filtering out false positives. This would limit its utility in exploratory analyses. JAFFA performed well with clean, simulated datasets but dropped in performance when presented with real datasets, picking up many false positives and missing true fusions. STAR-Fusion performed fairly well, with a fair sensitivity and fair specificity across data types. All of the comparison programs required a long time to run, ranging on the larger spike-in dataset from over 26 hours for MapSplice to 1 hour 18 minutes for the fastest, STAR-Fusion.

Pizzly showed high sensitivity across all datasets and high specificity in the simulated datasets. Presented with real data, pizzly returned fairly noisy results, with many false positives reported. In future work, we will apply additional filters to reduce these false positives while maintaining pizzly's high sensitivity. Pizzly runs much faster than any of the other programs, taking just 8.5 minutes on the large spike-in dataset. Since kallisto-pizzly runs in such a short amount of time, we are able to iteratively develop the pizzly filters based upon analysis of pizzly output and

biologically relevant heuristics. This should improve pizzly’s specificity. A discussion of these steps is described below in “Future steps.”

Table 3.2 75nt positive control

Pizzly detects the most fusions. The starred “false positives” reported by pizzly were filtered out in future pizzly iterations using our updated method discussed below. Timing is based on using 4 cores where possible.

tool	true positive	false positive	sensitivity %	clock time
MapSplice	42	1	84	2m1s
EricScript	39	0	78	9m34s
FusionCatcher	31	0	62	7m42s
JAFFA	43	0	86	3m56s
PRADA	32	0	64	4m5s
STAR-Fusion	45	1	90	6m59s
pizzly	46	2*	92	1m8s

Table 3.3 100nt positive control

tool	true positive	false positive	sensitivity %
MapSplice	39	1	78
EricScript	41	1	82
FusionCatcher	41	0	82
JAFFA	47	0	94
PRADA	0	0	0
STAR-Fusion	41	0	82
pizzly	46	0	92

Table 3.4 100nt negative control

All tools but EricScript perform well, picking up only minor numbers of false positives.

tool	false positive
MapSplice	3
EricScript	82
FusionCatcher	1
JAFFA	1
PRADA	1
STAR-Fusion	0
pizzly	1

Table 3.5 Spike-in real dataset containing nine true synthetic fusions

Since this is a real dataset, “false positives” has been replaced with “other” to allow for possible real fusions. In the study, only one other fusion was confirmed in the background RNA sample. Timing is based on using 4 cores where possible.

tool	true (9)	other	clock time
MapSplice	7	2	26h 37m
EricScript	9	575	10h 17m
FusionCatcher	9	3	7h 45m
JAFFA	7	458	20 17m
PRADA	0	0	19h 7m
STAR-Fusion	8	5	1h 18m
pizzly	9	322	8.5m

4. Future Steps

We are confident that pizzly false positives can be reduced with a few additional steps. For example, in the 75nt positive control, pizzly reported two false positives (starred in Table 3.2). Upon further investigation, we found that these false fusions originated from two transcripts of a gene overlapping the location of the true fusion gene. Pizzly predicts the “true” fusion BSG-COX6A1 and the “false” fusion BSG-AL021546.6 and identifies four possible transcript sequences arising from the two BSG transcripts and the two other transcripts. The original transcript estimates from kallisto predicted that only two of these sequences are present in the sample (Table 3.6). Given the breakpoint sequences of these four possible fusion transcripts, we can build a new kallisto index with the added potential fusions and re-quantify the original reads with kallisto using the updated index (Table 3.7).

Table 3.6 Original kallisto read counts for suspected fusion transcripts

transcript	estimated reads
COX6A1	3002
AL021546.6	0
BSG-002	2218
BSG-006	4

By adding a second kallisto step after fusion transcript assembly, we are able to increase detection of fusion transcripts while reducing false positives. In addition, this extra kallisto step allows us to estimate transcript fusion abundance. While all other fusion detection programs report transcript abundance only on the level of read counts, we are able to output the more meaningful and accurate transcript per million (TPM) report for detected fusions. To test the accuracy of TPM estimation for fusion transcripts, we applied the kallisto-pizzly-kallisto method to the spike-in dataset that contains fusion transcripts aliquoted at a known concentration. True fusion transcript

Table 3.7 Kallisto is able to correctly assign reads to fusions at the transcript level

transcript	estimated reads
COX6A1	0
AL021546.6	0
BSG-002	0
BSG-006	0
BSG-006 - AL021546.6	0
BSG-006 - COX6A1	0
BSG-002 - AL021546.6	5
BSG-002 - COX6A1	6139

abundance was calculated as follows: starting with 1ug of RNA, with mRNA being 1-5% of the total and each transcript being on average 2,500nt, the concentration of background RNA is approximately 1.17pmol and mRNA approximately 0.058pmol. The spike-in was aliquoted at -6.17 log10pmol. Under the assumption that the mRNA is between 1-5% of the total RNA, the total spike-in concentration should be 11-60 TPM divided across the nine synthetic fusions. Kallisto's estimated fusion abundances add up to 49.2TPM, which is within the expected range (Table 3.8).

Table 3.8 Kallisto reported TPM for the nine synthetic fusion transcripts

fusion	estimated counts	TPM
EWSR1-ATF1	310	4.46112
TMPRSS2-ETV1	602	7.57399
EWSR1-FLI1	296	3.12320
NTRK3-ETV6	829	6.37313
CD74-ROS1	843	5.77947
HOOK3-RET	986	6.06947
EML4-ALK	1798	7.42589
AKAP9-BRAF	568	1.92871
BRD4-NUT	1828	6.48389

With this additional step, pizzly becomes the first fusion detection tool to assign the industry-accepted transcript quantification measurement to its output fusions. Additionally, TPM estimates can be utilized to further reduce false positives. With this final step in place, we expect pizzly to perform as the optimal RNA-Seq based fusion detection, assembly, and quantification tool.

5. Supporting Information

This research was performed jointly with Páll Melsted, faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, with contributions from Isaac Charles Joseph, Graduate Program in Computational Biology, University of California, Berkeley, and Lior Pachter, Department of Mathematics, University of California, Berkeley. Pizzly was built upon kallisto, which was designed by Nicolas Bray, Innovative Genomics Initiative, University of California, Berkeley, Harold Pimentel, Department of Computer Science, University of California, Berkeley, Páll Melsted, and Lior Pachter. Public release of pizzly is planned for Winter 2016. Publication of pizzly methods and testing is planned for Spring 2017.

4 Circular RNA in *Drosophila melanogaster* embryos

I. Introduction

Circular RNA is a common and conserved class of RNA, exhibiting complex expression patterns in a variety of organisms. Circular transcripts result from the covalent bonding of the 3' and 5' termini of linear RNA. Researchers have long been aware of circular RNA (circRNA), but its abundance and role as a regulatory agent have just recently become apparent (see Figure 4.1 for a timeline of circular RNA discovery). In 1979 the first eukaryotic circular RNA was observed in HeLa cells via electron microscope (Hsu and Coca-Prados 1979). The novelty of the circular structure garnered some interest, but since traditional RNA purification techniques select for polyadenylated transcripts and discard the circularized molecules, the resulting paucity of examples to study offered little motivation for further interrogation of this peculiar RNA. Over the next decades, a small number of studies were conducted, some of which hinted at possible functional importance. These included the human RNA ANRIL circular isoform, which is correlated with atherosclerosis risk (Burd et al. 2010), and the mouse sex-determining gene SRY, which is first expressed as a linear isoform but circularizes at onset of puberty (Capel et al. 1993). It was only with the advent of high-throughput sequencing that researchers began to recognize that circRNA is manifest throughout biology. In 2011, two studies utilizing rRNA-depleted RNA-Seq uncovered thousands of circles in human and *Achaea* (Salzman et al. 2012, Danan et al. 2012). In 2013, the first evidence for functional circRNA was reported with the discovery of CDR1as, which is antisense to the cerebellar degeneration-related autoantigen 1 gene (CDR1), and acts as a miRNA sponge for miR-7 (Hansen et al. 2013).

Many circRNAs contain miRNA seed sites similar to that of CDR1as, though to a lesser extent. These circRNAs may regulate gene expression similarly to CDR1as. Other circRNAs not containing known miRNA seed sites could have other, as yet unknown functions. Notably, knockdown of CDR1as affected not only miR-7 targets, but also resulted in the downregulation of several housekeeping genes. This indicates that even CDR1as may have indirect or distinct roles from that of a miRNA sponge. In annotating human circRNA, Memczak et al. (2013) found that 85% of the circRNA transcripts they studied align sense to known genes. Most of these transcripts include between one and five exons, but only 65% utilize known splice sites at the circularization junction. This suggests that the standard splicing machinery are at work in the transcription and processing of circRNA, but that alternative regions are recognized in joining the 3' and 5' ends. An *in vitro* study by Pasma, Been, and Garcia-Blanco (1996) found that splicing of circRNA can be induced by inverted repeat regions flanking the desired circle. This was confirmed *in vivo* when Hansen et al. (2013) induced circular isoforms from a construct of the CDR1as exon flanked by inverted sequence in HeLa Flp-in cells. Jeck et al. (2013) found that ALU

repeats are associated with splicing of circRNA in mammals. These studies indicate that flanking repeat elements are sufficient to induce circRNA within gene regions. Additional factors must also contribute to circRNA expression, as 10% of the circRNA in the study by Memcsak et al. (2013) aligned antisense to known genes, and the remaining 5% aligned to introns, UTRs and other genomic regions that do not rely on mRNA transcription.

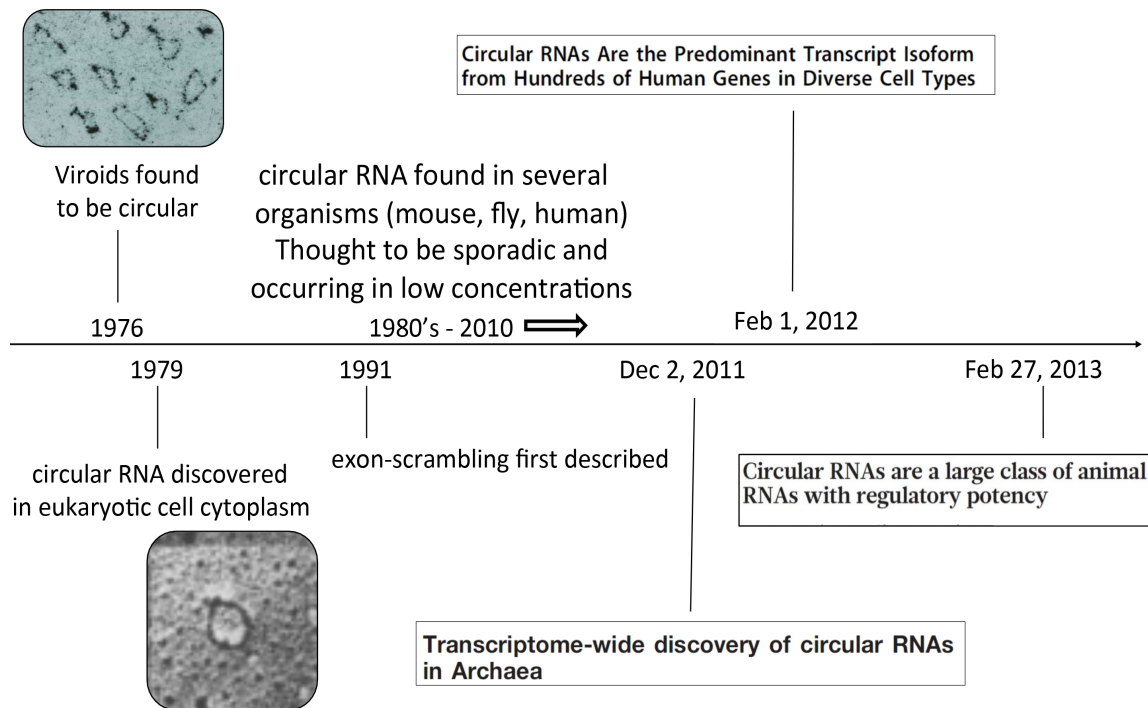


Figure 4.1 Circular RNA research timeline

Circular RNA was thought to be sporadic and in low concentrations until around 2011 when it was discovered that circular RNA is common and functional. The top left image is an electron micrograph of potato spindle tuber viroids. Lower left image depicts an electron micrograph of circular RNA extracted from HeLa cell cytoplasm (Hsu and Coca-Prados 1979). The boxed items are publication titles from the dates indicated.

Once formed, circRNA are concentrated in the cytosol (Salzman et al. 2012). Transport of circRNA to the cytosol is undefined. Degradation of circRNA is also poorly understood (Hansen et al. 2011), though it has been shown that circRNA degrades slower *in vivo* than linear transcripts (Memczak et al. 2013). There are many conjectured roles for circRNA: it may bind to 3' UTRs to regulate target mRNA expression, or it might help assemble larger molecular complexes. Since circular RNAs are more stable than their linear counterparts, they could be circularizing as a means of long-term storage, to linearize and translate as needed. To clearly ascertain circular RNA's biological role, it is necessary to understand its production and regulation. An effective method for this is to examine circular RNA in RNA-Seq data. To this end, we utilized RNA-Seq of ribosomal RNA-depleted total RNA to *de*

de novo annotate circular RNA within *D. melanogaster* embryos across six developmental stages using a novel computational method for circular read detection.

2. Results and Discussion

To evaluate circRNA in *Drosophila melanogaster*, we employed a combination of computational and molecular methods. We used a modified RNA-Seq library preparation strategy to increase the percentage of circular transcripts in our samples, and developed a computational workflow for circRNA transcript detection to *de novo* annotate circRNA within *D. melanogaster*.

2.1 A Test of Circular RNA Sample Purification Methods

CircRNA is not polyadenylated, so standard poly(A+) purification for sequencing preparation is not an option. We sought to find the best method to concentrate circRNA within a total RNA sample. One strategy we employed was to treat the samples with exoribonuclease to degrade linear transcripts, leaving circular and other high structure transcripts behind. The other strategy we applied was to selectively remove ribosomal RNA, which makes up about 85% of total RNA. To test these two strategies, we extracted total RNA from pooled stage 5-6 embryos and separated it into six samples. After a preliminary treatment with DNase, the following treatments were applied to the total RNA samples: Two were treated with Ribo-Zero (R0) rRNA removal kit according to instructions (Sooknanan, Pease, and Doyle 2010), which is designed for human/rat/mouse rRNA, but works fairly well for *Drosophila*, leaving the 5S rRNA (Epibio 2010); two were treated with RNase R (Suzuki and Tsukahara 2014), a 3' to 5' exoribonuclease which selectively degrades RNA with a 3' overhang of 7nt or greater. RNase R has difficulties degrading tRNA and the 5S rRNA, due to the structure of these molecules. The remaining two samples were treated with both R0 and RNase R. The RNase R samples should be enriched with circRNA and devoid of linear counterparts and we expect the same circRNA to be detected in RNase R treated and untreated sets, with enrichment for rare circles in the treated set. After treatment, the samples were run on the bioanalyzer, which showed similar levels of rRNA depletion and peaks of RNA concentration around 100nt. The samples were converted into sequencing libraries using standard Illumina protocol, and 50nt single-end reads were sequenced on the HiSeq 2000. Reads were then aligned to the reference transcriptome, with ribosomal sequence added, using Bowtie. Table 4.1 and Figure 4.2 depict the results of this study. RNase R did not remove much rRNA, likely because of the complex structure of these transcripts. When R0 was used in conjunction with RNase R, we see that a larger percentage of the reads aligns to rRNA than when R0 is used alone. This is most likely due to the degradation of other transcripts in the sample, concentrating the remaining transcripts. These datasets would be useful in further comparisons with other circular RNA datasets and

in assessing the level of structure of various transcripts. As a proof-of-concept to check for presence of circular RNA in the samples before proceeding to a more detailed sample collection, we created a reference of all possible within-gene exon junctions and aligned each sample to it to check for evidence of circular reads. We found multiple circular RNA junction reads. For instance, sample 5 included, among many others, the junctions (listed as gene 5'exon_3'exon): exba 13_9, Rox8 9_5, Uhg3 3_3, uex 7_12, unc-13 9_4, larp 9_9, and sdt 17_14. For this study, this is all of the analysis that was done on these pooled samples.

Table 4.1 Ribosomal reads remaining post treatment

Percentage of total reads mapping to ribosomal transcripts after rRNA removal and/or RNase treatment.

sample	2	4	5	6	7	12
treatment	R0	R0	R0+RNase R	R0+RNase R	RNase R	RNase R
total reads	32,937,214	30,938,042	34,300,687	37,654,707	34,734,808	43,142,999
reads aligned	19,257,093	22,161,528	18,836,709	20,222,805	24,327,916	31,938,320
percent aligned	58.5%	71.6%	54.9%	53.7%	70.0%	74.0%
percent ribosomal	8.6%	11.3%	25.5%	23.8%	66.8%	72.1%
18S	2.6%	1.3%	0.5%	1.2%	10.7%	6.3%
28S	87.2%	62.2%	16.2%	18.9%	21.2%	11.1%
5.8S	0.308%	4.719%	2.402%	2.309%	0.022%	0.021%
5S	9.9%	31.8%	80.9%	77.6%	68.1%	82.5%

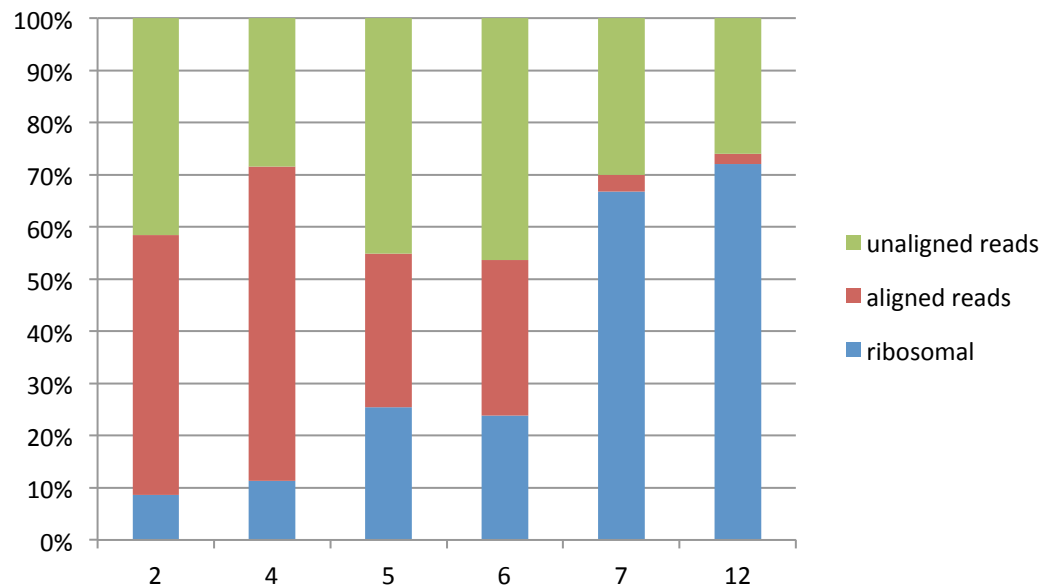


Figure 4.2 Division of reads from samples treated with RNase or rRNA depletion

The red “aligned reads” are all reads mapping to transcripts that are not ribosomal. Samples 5 and 6 show a depletion of rRNA and an increase of unaligned reads, indicating that these samples removed linear transcripts and rRNA while concentrating for non-coding RNA, and potential circRNA junction reads.

2.2 Staged Single Embryo Collection and Sequencing

Wild type Oregon R flies were grown under standard laboratory conditions and collected at six embryonic time points, as shown in Figure 4.3. For collection, eggs were incubated until specified stage, dechorionated using a 50% bleach solution, sorted and staged under microscope, and briefly stored on ice until RNA was extracted using TRIzol reagent. A range of time points allows for differential expression analysis of circRNA throughout embryonic development. These six time points specifically were chosen based upon their easy-to-discern phenotype as well as distribution across post-cellularization embryonic development (Campos-Ortega and Hartenstein 1985). Total RNA was extracted from two single embryos at each stage and treated with Ribo-Zero. Strand-specific, multiplexed libraries were prepared using the Illumina RNA TruSeq protocol and sequenced on one lane of an Illumina HiSeq 2500 as 100nt paired-end reads. We only obtained one sample for the stage, “trachea fills with air.” Reads were mapped to *D. melanogaster* reference transcriptome using eXpress. PCA plot of transcripts per million (TPM) for samples (Figure 4.4) shows biological replicates group together and time is correlated with principal component one. Remarkably, samples associate along PCA1 in chronological order.

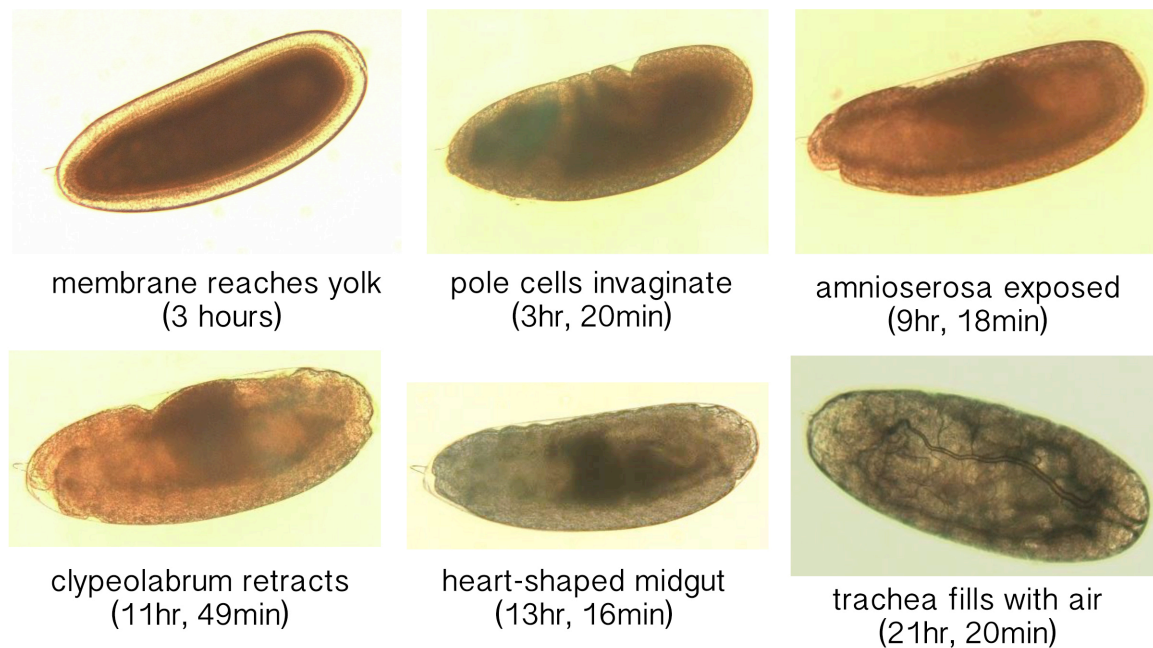


Figure 4.3 *D. melanogaster* embryos at six developmental stages

Single embryos were collected at six developmental stages. Pictured here are one of each of the embryos that were collected and sequenced. Stage descriptions are associated with the following standard stage numbers: membrane reaches yolk: 5, pole cells invaginate: 7, amnioserosa exposed: 12, clypeolabrum retracts: 14, heart-shaped midgut: 15, trachea fills with air: 17.

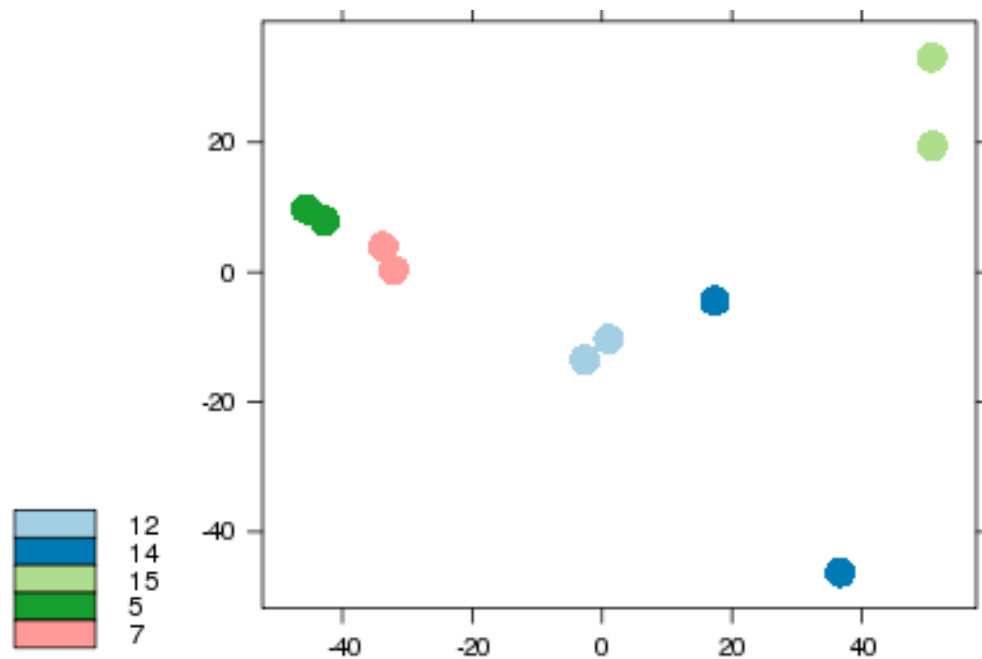


Figure 4.4 Principal component analysis plot of five embryo stages

Plot shows that replicates group together, and principal component one is correlated with developmental stage. Interestingly, the samples span PCA1 in chronological order. Stage “trachea fills with air” was not included in plot as it was too far to the right and disturbed the scale of the other samples. Legend gives stage number as described in Figure 4.3.

2.3 Computational Workflow for Circular RNA Detection

In RNA-Seq, paired-end reads from linear transcripts map with the first read of the fragment 5’ of and pointing towards the downstream second read. For circular RNA, fragments flanking the 3’-5’ splice junction map to the reference in the opposite order, with the first read 3’ of and pointing away from the second read of the fragment (Figure 4.5). As a result, standard RNA-Seq alignment methods discard reads from circular junctions. We can take advantage of this mapping distinction to select for circRNA fragments with reads flanking the circular splice junction.

Existing circRNA detection methods require several compromises to work correctly, including (1) the requirement of junction-spanning reads where a single read is split between the two parts of the circular junction, (2) biased mapping to specific junction reference sets, e.g. an exon-junction custom reference, and (3) the inability to determine the full sequence of the circular transcripts (Danan et al. 2012; Salzman et al. 2013; Jeck et al. 2013). The first two issues we can resolve by a stepwise mapping that first detects the flanking reads, as described above, and then detects spanning reads as performed in existing methods. The third issue of quantifying the circular transcripts is a more nuanced challenge. Existing workflows can only define circRNA near tail-to-head junction regions. Any reads that map ambiguously to the intervening sequence in candidate circles and to linear transcripts of the same genomic region are currently unassigned. Intervening sequence can be

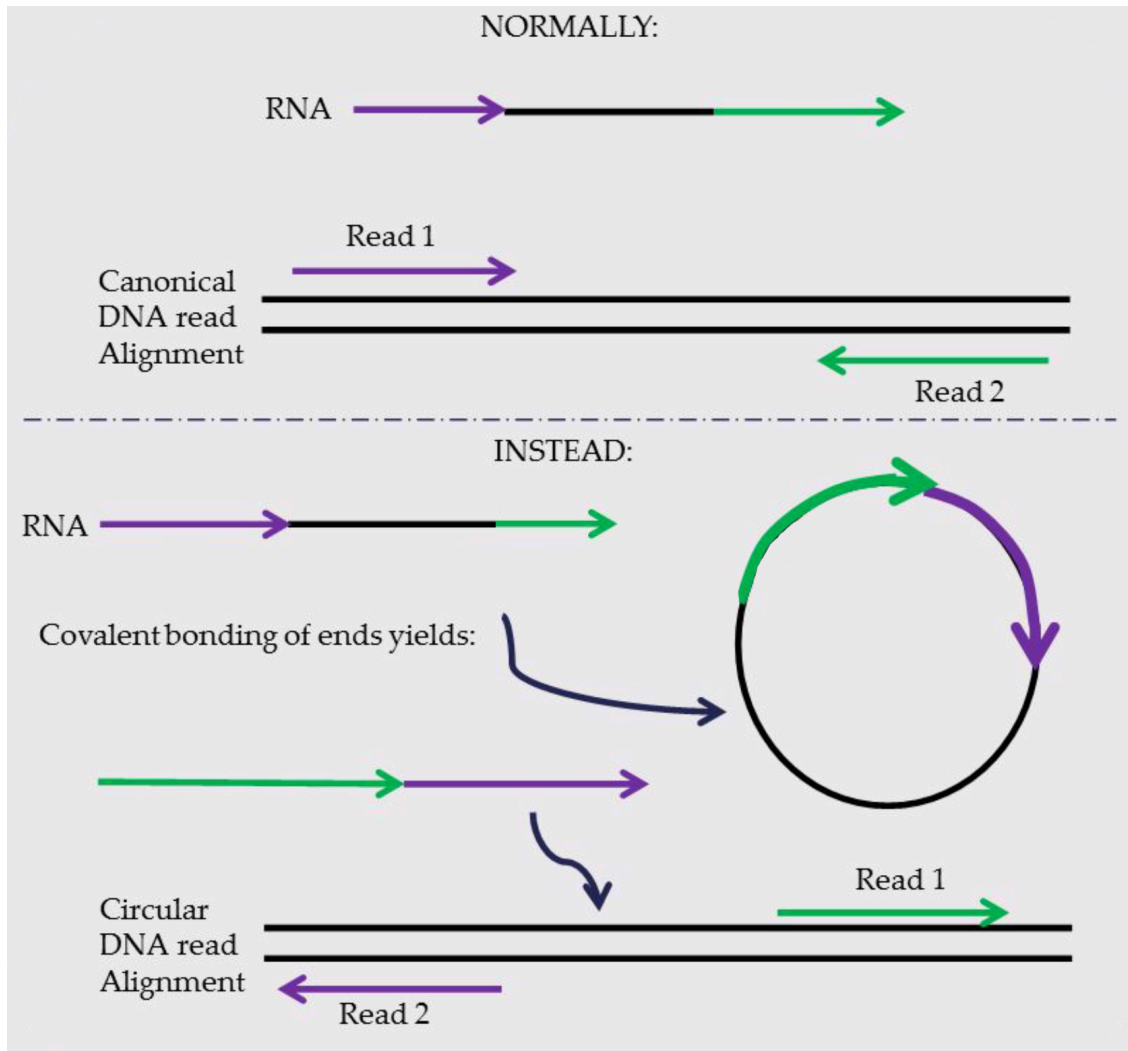


Figure 4.5 Circular RNA fragment mapping in RNA-Seq

Normally, transcripts proceed in 5' to 3' direction and RNA-Seq reads point towards each other when mapped to a reference. In circular RNA, the downstream donor site covalently bonds to upstream acceptor site. Fragmentation yields 3' upstream of 5'. Hence, reads point away from each other.

determined through molecular interrogation, but this is not scalable to large datasets. Accurate detection and quantification of circRNA transcripts requires that these intervening reads be disambiguated. To solve this problem, the mapped circRNA reads are submitted to two additional steps that accurately assign the intervening reads to their corresponding transcripts: (1) use the mapped circRNA reads to create a circRNA transcriptome, and (2) quantify transcript abundance of these reads using a modified version of the eXpress transcript quantification software (Roberts and Pachter 2013). The custom circRNA detection and quantification workflow is described in Figure 4.6 and its caption.

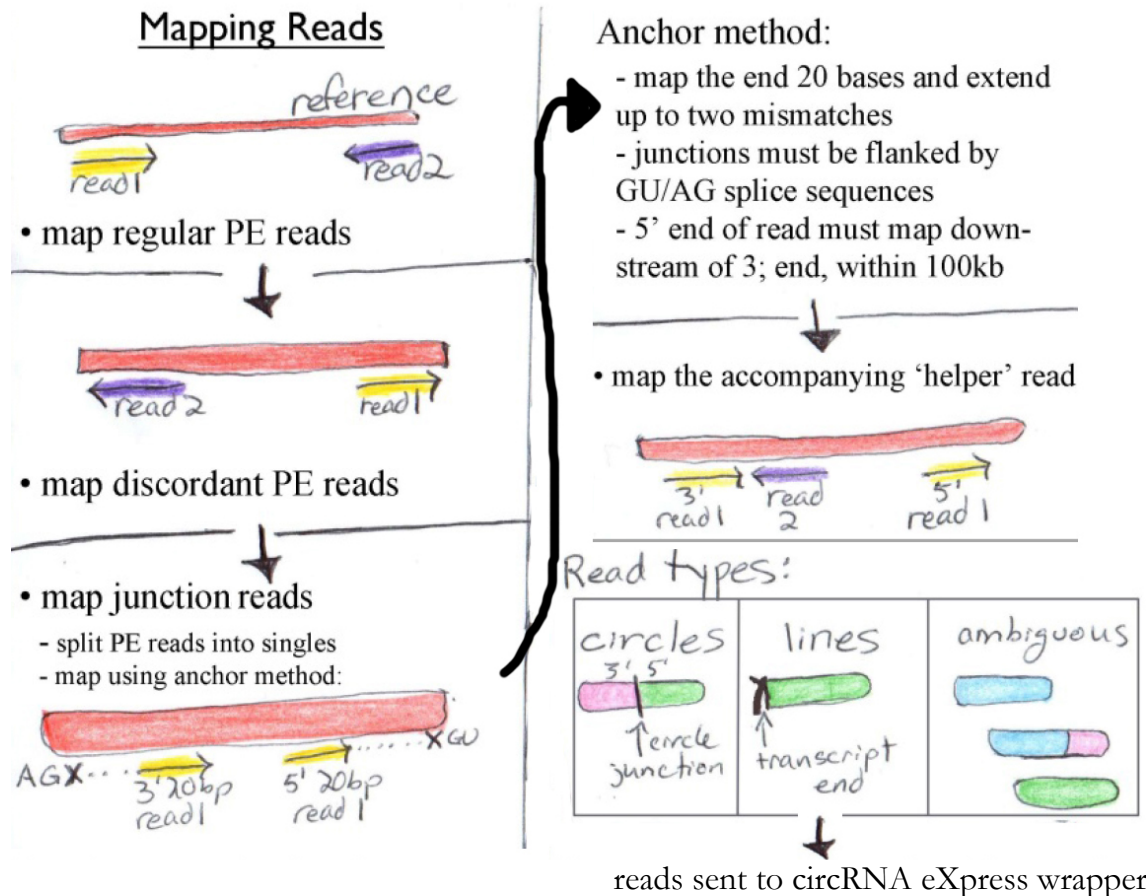


Figure 4.6 Workflow for circRNA mapping and transcript assignment

Paired end reads are first mapped to the basic reference transcriptome. Reads are mapped, and the orientation, either regular or discordant, is noted. If a read does not map, we then try mapping each end of the read using the 'anchor method' described by others previously. All read fragments can be assigned to one of three types: circles - which cover a 3'-5' junction, linear transcripts - which contain a transcript end, and ambiguous reads - which could come from either transcript type. A custom transcriptome is constructed using this read information that includes the candidate circular transcript sequences. Reads are then run through a circRNA eXpress wrapper to determine the complete sequences and abundance of the linear and circular transcripts.

3. Future Directions

This work on circular RNA was undertaken over the years 2012-2013. Since then, members of the Pachter lab have developed the RNA-Seq quantification software, kallisto (Bray et al. 2016), and its accompanying transcript fusion detection tool, pizzly. With a few modifications, pizzly could be converted into a tool for fast and accurate circular RNA detection. Even now, in 2016, there is a major need for better RNA-Seq circRNA analysis tools. There are five major circRNA detection algorithms currently in use: circRNA finder, find_circ, CIRCexplorer, CIRI, and MapSplice (Hansen et al. 2015). All give very different predictions. As such, the best method at the moment for circRNA detection is to compare output from multiple tools and take the overlapping circles as "truth." Kallisto provides a faster and more accurate

solution to transcript assignment and quantification. These advantages should carry over to circRNA detection as well. The general idea takes into consideration that kallisto pseudoaligns k-mers to reference transcripts, skipping from one informative k-mer to the next. When an error is encountered in the next k-mer, kallisto goes back to all intermediate k-mers and checks them for concordance with previous k-mers. This step is where we can program kallisto to search for circular transcripts. If the read originated from the junction of a circular transcript, all intervening k-mers up to the point of the junction will be concordant with the previous k-mer. At the point of “mismapping,” i.e. the circular junction, the intervening k-mers will “map” to the reference in concordance with the k-mer that signaled the error in the first place. Such reads will be partitioned to a separate pseudo-bam file as candidate circular junctions. Junction position can be recorded and checked for 3'-5' orientation.

The next steps in the development of a robust circRNA detection tool would be to (1) modify kallisto for circRNA detection, (2) test and benchmark the tool using simulated datasets, and (3) to finally analyze the circular RNA in our staged, single-embryo *D. melanogaster* samples.

Bibliography

- Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. *Molecular Biology of the Cell*. 4th ed. Garland Science.
- Barrell, B. G., A. T. Bankier, and J. Drouin. 1979. "A Different Genetic Code in Human Mitochondria." *Nature* 282 (5735): 189–94. doi:10.1038/282189a0.
- Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity." *Nature* 483 (7391): 603–7. doi:10.1038/nature11003.
- Bar-Yaacov, Dan, Gal Avital, Liron Levin, Allison Richards, Naomi Hachen, Boris Rebolledo Jaramillo, Anton Nekrutenko, Raz Zarivach, and Dan Mishmar. 2013. "RNA-DNA Differences in Human Mitochondria Restore Ancestral Form of 16S Ribosomal RNA." *Genome Research*, August. doi:10.1101/gr.161265.113.
- Beaumeunier, S., J. Audoux, A. Boureux, T. Commes, N. Philippe, and R. Alves. 2015. "The Role of Machine Learning in Finding Chimeric RNAs." In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, 26–30. doi:10.1109/DEXA.2015.25.
- Böck, A., K. Forchhammer, J. Heider, W. Leinfelder, G. Sawers, B. Veprek, and F. Zinoni. 1991. "Selenocysteine: The 21st Amino Acid." *Molecular Microbiology* 5 (3): 515–20. doi:10.1111/j.1365-2958.1991.tb00722.x.
- Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics*, April, btu170. doi:10.1093/bioinformatics/btu170.
- Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016a. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* advance online publication (April). doi:10.1038/nbt.3519.
- Brown, James B., Nathan Boley, Robert Eisman, Gemma E. May, Marcus H. Stoiber, Michael O. Duff, Ben W. Booth, et al. 2014. "Diversity and Dynamics of the *Drosophila* Transcriptome." *Nature* advance online publication (March). doi:10.1038/nature12962.
- Burd, Christin E., William R. Jeck, Yan Liu, Hanna K. Sanoff, Zefeng Wang, and Norman E. Sharpless. 2010. "Expression of Linear and Novel Circular Forms of an INK4/ARF-Associated Non-Coding RNA Correlates with Atherosclerosis Risk." *PLoS Genetics* 6 (12). doi:10.1371/journal.pgen.1001233.
- Burighel, P., F. Caicci, and L. Manni. 2011. "Hair Cells in Non-Vertebrate Models: Lower Chordates and Molluscs." *Hearing Research*, Comparative Studies of the Ear, 273 (1–2): 14–24. doi:10.1016/j.heares.2010.03.087.
- Cadenas, Enrique, and Kelvin J. A. Davies. 2000. "Mitochondrial Free Radical Generation, Oxidative Stress, and aging1." *Free Radical Biology and Medicine* 29 (3–4): 222–30. doi:10.1016/S0891-5849(00)00317-8.

- Campos-Ortega, José A., and Volker Hartenstein. 1985. "Stages of *Drosophila* Embryogenesis." In *The Embryonic Development of Drosophila Melanogaster*, 9–84. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-662-02454-6_3.
- Cancer Genome Atlas Research Network, Daniel J. Brat, Roel G. W. Verhaak, Kenneth D. Aldape, W. K. Alfred Yung, Sofie R. Salama, Lee A. D. Cooper, et al. 2015. "Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas." *The New England Journal of Medicine* 372 (26): 2481–98. doi:10.1056/NEJMoa1402121.
- Capel, B, A Swain, S Nicolis, A Hacker, M Walter, P Koopman, P Goodfellow, and R Lovell-Badge. 1993. "Circular Transcripts of the Testis-Determining Gene Sry in Adult Mouse Testis." *Cell* 73 (5): 1019–30.
- Cheng Zhu, Gang Bao, and Ning Wang. 2000. "Cell Mechanics: Mechanical Response, Cell Adhesion, and Molecular Deformation." *Annual Review of Biomedical Engineering* 2 (1): 189–226. doi:10.1146/annurev.bioeng.2.1.189.
- Chin, Lynda, Jannik N. Andersen, and P. Andrew Futreal. 2011. "Cancer Genomics: From Discovery Science to Personalized Medicine." *Nature Medicine* 17 (3): 297–303. doi:10.1038/nm.2323.
- Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szczesniak, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." *Genome Biology* 17: 13. doi:10.1186/s13059-016-0881-8.
- Crucian, Brian, Raymond Stowe, Satish Mehta, Peter Uchakin, Heather Quiriarte, Duane Pierson, and Clarence Sams. 2012. "Immune System Dysregulation Occurs During Short Duration Spaceflight On Board the Space Shuttle." *Journal of Clinical Immunology* 33 (2): 456–65. doi:10.1007/s10875-012-9824-7.
- Danan, Miri, Schraga Schwartz, Sarit Edelheit, and Rotem Sorek. 2012. "Transcriptome-Wide Discovery of Circular RNAs in Archaea." *Nucleic Acids Research* 40 (7): 3131–42. doi:10.1093/nar/gkr1009.
- Davies, Shireen A., Pablo Cabrero, Gayle Overend, Lorraine Aitchison, Sujith Sebastian, Selim Terhzaz, and Julian A. T. Dow. 2014. "Cell Signalling Mechanisms for Insect Stress Tolerance." *Journal of Experimental Biology* 217 (1): 119–28. doi:10.1242/jeb.090571.
- Doctor, John, Dianne Fristrom, and James Fristrom. 1985. "The Pupal Cuticle of *Drosophila*: Biphaseic Synthesis of Pupal Cuticle Proteins in Vivo and in Vitro in Response to 20-Hydroxyecdysone." *The Journal of Cell Biology* 101 (1): 189–200.
- Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30 (1): 207–10. doi:10.1093/nar/30.1.207.
- Epibio. 2010. "rRNA Removal from *Drosophila* Total RNA." *EpiCentral*. October 27. <http://epicentral.blogspot.com/2010/10/rna-removal-from-drosophila-total-rna.html>.

- Feder, Martin E., and Gretchen E. Hofmann. 1999. "HEAT-SHOCK PROTEINS, MOLECULAR CHAPERONES, AND THE STRESS RESPONSE: Evolutionary and Ecological Physiology." *Annual Review of Physiology* 61 (1): 243–82. doi:10.1146/annurev.physiol.61.1.243.
- Flatt, Thomas, Andreas Heyland, Florentina Rus, Ermelinda Porpiglia, Chris Sherlock, Rochele Yamamoto, Alina Garbuzov, Subba R. Palli, Marc Tatar, and Neal Silverman. 2008. "Hormonal Regulation of the Humoral Innate Immune Response in *Drosophila Melanogaster*." *Journal of Experimental Biology* 211 (16): 2712–24. doi:10.1242/jeb.014878.
- Ge, Huanying, Kejun Liu, Todd Juan, Fang Fang, Matthew Newman, and Wolfgang Hoeck. 2011. "FusionMap: Detecting Fusion Genes from next-Generation Sequencing Data at Base-Pair Resolution." *Bioinformatics* 27 (14): 1922–28. doi:10.1093/bioinformatics/btr310.
- Gingeras, Thomas R. 2009. "Implications of Chimaeric Non-Co-Linear Transcripts." *Nature* 461 (7261): 206–11.
- Girardot, Fabrice, Véronique Monnier, and Hervé Tricoire. 2004. "Genome Wide Analysis of Common and Specific Stress Responses in Adult *Drosophila Melanogaster*." *BMC Genomics* 5 (1): 74. doi:10.1186/1471-2164-5-74.
- Goldermann, Markus, and Wolfgang Hanke. 2001. "Ion Channel Are Sensitive to Gravity Changes." *Microgravity Science and Technology* 13 (1): 35–38. doi:10.1007/BF02873330.
- Guéguinou, Nathan, Mickaël Bojados, Marc Jamon, Hanane Derradji, Sarah Baatout, Eric Tschirhart, Jean-Pol Fripiat, and Christine Legrand-Frossi. 2012. "Stress Response and Humoral Immune System Alterations Related to Chronic Hypergravity in Mice." *Psychoneuroendocrinology* 37 (1): 137–47. doi:10.1016/j.psyneuen.2011.05.015.
- Hanahan, Douglas, and Robert A Weinberg. 2000. "The Hallmarks of Cancer." *Cell* 100 (1): 57–70. doi:10.1016/S0092-8674(00)81683-9.
- Hansen, Thomas B., Trine I. Jensen, Bettina H. Clausen, Jesper B. Bramsen, Bente Finsen, Christian K. Damgaard, and Jørgen Kjems. 2013. "Natural RNA Circles Function as Efficient microRNA Sponges." *Nature* 495 (7441): 384–88. doi:10.1038/nature11993.
- Hansen, Thomas B., Morten T. Venø, Christian K. Damgaard, and Jørgen Kjems. 2015. "Comparison of Circular RNA Prediction Tools." *Nucleic Acids Research*, benchmarking, , December, gkv1458. doi:10.1093/nar/gkv1458.
- Hansen, Thomas B., Erik D. Wiklund, Jesper B. Bramsen, Sune B. Villadsen, Aaron L. Statham, Susan J. Clark, and Jørgen Kjems. 2011. "miRNA-dependent Gene Silencing Involving Ago2-mediated Cleavage of a Circular Antisense RNA." *The EMBO Journal* 30 (21): 4414–22. doi:10.1038/emboj.2011.359.
- Han, Yixing, Shouguo Gao, Kathrin Muegge, Wei Zhang, and Bing Zhou. 2015. "Advanced Applications of RNA Sequencing and Challenges." *Bioinformatics and Biology Insights* 9 (Suppl 1): 29–46. doi:10.4137/BBI.S28991.
- Hateley, Shannon, Ravikumar Hosamani, Shilpa R. Bhardwaj, Lior Pachter, and Sharmila Bhattacharya. 2016. "Transcriptomic Response of *Drosophila*

- Melanogaster Pupae Developed in Hypergravity.” *Genomics* 108 (3–4): 158–67. doi:10.1016/j.ygeno.2016.09.002.
- Herranz, Raul, Alberto Benguría, David A. Laván, Irene López-Vidriero, Gilbert Gasset, F. Javier Medina, Jack J. W. A. Van LOON, and Roberto Marco. 2010. “Spaceflight-Related Suboptimal Conditions Can Accentuate the Altered Gravity Response of *Drosophila* Transcriptome.” *Molecular Ecology* 19 (19): 4255–64. doi:10.1111/j.1365-294X.2010.04795.x.
- Herranz, Raul, Oliver J. Larkin, Camelia E. Dijkstra, Richard JA Hill, Paul Anthony, Michael R. Davey, Laurence Eaves, Jack JWA van Loon, F. Javier Medina, and Roberto Marco. 2012. “Microgravity Simulation by Diamagnetic Levitation: Effects of a Strong Gradient Magnetic Field on the Transcriptional Profile of *Drosophila* Melanogaster.” *BMC Genomics* 13: 52. doi:10.1186/1471-2164-13-52.
- Herranz, Raul, Oliver J. Larkin, Richard JA Hill, Irene Lopez-Vidriero, Jack JWA van Loon, and F. J. Medina. 2013. “Suboptimal Evolutionary Novel Environments Promote Singular Altered Gravity Responses of Transcriptome during *Drosophila* Metamorphosis.” *BMC Evolutionary Biology* 13 (1): 133. doi:10.1186/1471-2148-13-133.
- Holoch, Daniel, and Danesh Moazed. 2015. “RNA-Mediated Epigenetic Regulation of Gene Expression.” *Nature Reviews Genetics* 16 (2): 71–84. doi:10.1038/nrg3863.
- Hsu, M T, and M Coca-Prados. 1979. “Electron Microscopic Evidence for the Circular Form of RNA in the Cytoplasm of Eukaryotic Cells.” *Nature* 280 (5720): 339–40.
- Jander, Rudolf. 1963. “Insect Orientation.” *Annual Review of Entomology* 8 (1): 95–114. doi:10.1146/annurev.en.08.010163.000523.
- Jeck, William R., Jessica A. Sorrentino, Kai Wang, Michael K. Slevin, Christin E. Burd, Jinze Liu, William F. Marzluff, and Norman E. Sharpless. 2013. “Circular RNAs Are Abundant, Conserved, and Associated with ALU Repeats.” *RNA* 19 (2): 141–57. doi:10.1261/rna.035667.112.
- Kang, Kyung Shin, Seung-Jae Lee, Haksue Lee, Wonkyu Moon, and Dong-Woo Cho. 2011. “Effects of Combined Mechanical Stimulation on the Proliferation and Differentiation of Pre-Osteoblasts.” *Experimental & Molecular Medicine* 43 (6): 367–73. doi:10.3858/emm.2011.43.6.040.
- Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. “TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions.” *Genome Biology* 14 (4): R36. doi:10.1186/gb-2013-14-4-r36.
- Kolesnikov, Yaroslav S., Serhiy V. Kretynin, Igor D. Volotovskiy, Elizabeth L. Kordyum, Eric Ruelland, and Volodymyr S. Kravets. 2015. “Molecular Mechanisms of Gravity Perception and Signal Transduction in Plants.” *Protoplasma*, July, 1–18. doi:10.1007/s00709-015-0859-5.
- Kruger, Kelly, Paula J. Grabowski, Arthur J. Zaug, Julie Sands, Daniel E. Gottschling, and Thomas R. Cech. 1982. “Self-Splicing RNA: Autoexcision and

- Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena.” *Cell* 31 (1): 147–57. doi:10.1016/0092-8674(82)90414-7.
- Kumar, Shailesh, Sundus Khalid Razzaq, Angie Duy Vo, Mamta Gautam, and Hui Li. 2016. “Identifying Fusion Transcripts Using next Generation Sequencing.” *Wiley Interdisciplinary Reviews: RNA*, August, n/a – n/a. doi:10.1002/wrna.1382.
- Kumar, Shailesh, Angie Duy Vo, Fujun Qin, and Hui Li. 2016. “Comparative Assessment of Methods for the Fusion Transcripts Detection from RNA-Seq Data.” *Scientific Reports* 6 (February): 21597. doi:10.1038/srep21597.
- Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. doi:10.1038/nmeth.1923.
- Lebo, Matthew S, Laura E Sanders, Fengzhu Sun, and Michelle N Arbeitman. 2009. “Somatic, Germline and Sex Hierarchy Regulated Gene Expression during Drosophila Metamorphosis.” *BMC Genomics* 10 (February): 80. doi:10.1186/1471-2164-10-80.
- Le Bourg, Éric. 2012. “Combined Effects of Two Mild Stresses (cold and Hypergravity) on Longevity, Behavioral Aging, and Resistance to Severe Stresses in Drosophila Melanogaster.” *Biogerontology* 13 (3): 313–28. doi:10.1007/s10522-012-9377-4.
- Le Bourg, Éric, and Didier Fournier. 2004. “Is Lifespan Extension Accompanied by Improved Antioxidant Defences? A Study of Superoxide Dismutase and Catalase in Drosophila Melanogaster Flies That Lived in Hypergravity at a Young Age.” *Biogerontology* 5 (4): 261–66. doi:10.1023/B:BGEN.0000038046.37590.03.
- Le Bourg, Éric, Nadège Minois, Pol Bullens, and Philippe Bearet. 2000. “A Mild Stress due to Hypergravity Exposure at Youngage Increases Longevity in Drosophila Melanogaster Males.” *Biogerontology* 1 (2): 145–55. doi:10.1023/A:1010043914016.
- Lee, S.-I., K.-H. Park, S.-J. Kim, Y.-G. Kang, Y.-M. Lee, and E.-C. Kim. 2012. “Mechanical Stress-Activated Immune Response Genes via Sirtuin 1 Expression in Human Periodontal Ligament Cells.” *Clinical & Experimental Immunology* 168 (1): 113–24. doi:10.1111/j.1365-2249.2011.04549.x.
- Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12: 323. doi:10.1186/1471-2105-12-323.
- Lightowlers, Robert N., Agata Rozanska, and Zofia M. Chrzanowska-Lightowlers. 2014. “Mitochondrial Protein Synthesis: Figuring the Fundamentals, Complexities and Complications, of Mammalian Mitochondrial Translation.” *FEBS Letters* 588 (15): 2496–2503. doi:10.1016/j.febslet.2014.05.054.
- Li, Hui, Jinglan Wang, Xianrong Ma, and Jeffrey Sklar. 2009. “Gene Fusions and RNA Trans-Splicing in Normal and Neoplastic Human Cells.” *Cell Cycle* 8 (2): 218–22.
- Liu, Silvia, Wei-Hsiang Tsai, Ying Ding, Rui Chen, Zhou Fang, Zhiguang Huo, SungHwan Kim, et al. 2016. “Comprehensive Evaluation of Fusion Transcript Detection Algorithms and a Meta-Caller to Combine Top Performing

- Methods in Paired-End RNA-Seq Data.” *Nucleic Acids Research* 44 (5): e47–e47. doi:10.1093/nar/gkv1234.
- Lodish, Harvey, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. 2000. “Processing of rRNA and tRNA.” <https://www.ncbi.nlm.nih.gov/books/NBK21729/>.
- Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550. doi:10.1186/s13059-014-0550-8.
- Marcu, Oana, Matthew P. Lera, Max E. Sanchez, Edina Levic, Laura A. Higgins, Alena Shmygelska, Thomas F. Fahlen, Helen Nichol, and Sharmila Bhattacharya. 2011. “Innate Immune Responses of *Drosophila Melanogaster* Are Altered by Spaceflight.” *PLOS ONE* 6 (1): e15361. doi:10.1371/journal.pone.0015361.
- Memczak, Sebastian, Marvin Jens, Antignoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, et al. 2013. “Circular RNAs Are a Large Class of Animal RNAs with Regulatory Potency.” *Nature* 495 (7441): 333–38. doi:10.1038/nature11928.
- Mertens, Fredrik, Bertil Johansson, Thoas Fioretos, and Felix Mitelman. 2015. “The Emerging Complexity of Gene Fusions in Cancer.” *Nature Reviews Cancer* 15 (6): 371–81. doi:10.1038/nrc3947.
- Mi, Huaiyu, Sagar Poudel, Anushya Muruganujan, John T. Casagrande, and Paul D. Thomas. 2016. “PANTHER Version 10: Expanded Protein Families and Functions, and Analysis Tools.” *Nucleic Acids Research* 44 (D1): D336–42. doi:10.1093/nar/gkv1194.
- Minois, Nadège. 2006. “The Hormetic Effects of Hypergravity on Longevity and Aging.” *Dose-Response* 4 (2): 145–54. doi:10.2203/dose-response.05-008.Minois.
- Mitsui, Y., R. Langridge, B. E. Shortle, C. R. Cantor, R. C. Grant, M. Kodama, and R. D. Wells. 1970. “Physical and Enzymatic Studies on Poly d(I-C)-Poly d(I-C), an Unusual Double-Helical DNA.” *Nature* 228 (5277): 1166–69.
- Morris, Kevin V., and John S. Mattick. 2014. “The Rise of Regulatory RNA.” *Nature Reviews Genetics* 15 (6): 423–37. doi:10.1038/nrg3722.
- Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. doi:10.1038/nmeth.1226.
- Moussian, Bernard, Heinz Schwarz, Slawomir Bartoszewski, and Christiane Nüsslein-Volhard. 2005. “Involvement of Chitin in Exoskeleton Morphogenesis in *Drosophila Melanogaster*.” *Journal of Morphology* 264 (1): 117–30. doi:10.1002/jmor.10324.
- Nissen, Poul, Jeffrey Hansen, Nenad Ban, Peter B. Moore, and Thomas A. Steitz. 2000. “The Structural Basis of Ribosome Activity in Peptide Bond Synthesis.” *Science* 289 (5481): 920–30. doi:10.1126/science.289.5481.920.
- “Ontology Structure | Gene Ontology Consortium.” 2016. Accessed July 29. <http://geneontology.org/page/ontology-structure>.

- Ozsolak, Fatih, and Patrice M Milos. 2011. "RNA Sequencing: Advances, Challenges and Opportunities." *Nature Reviews. Genetics* 12 (2): 87–98. doi:10.1038/nrg2934.
- Ozsolak, Fatih, Adam R. Platt, Dan R. Jones, Jeffrey G. Reifengerger, Lauryn E. Sass, Peter McInerney, John F. Thompson, Jayson Bowers, Mirna Jarosz, and Patrice M. Milos. 2009. "Direct RNA Sequencing." *Nature* 461 (7265): 814–18. doi:10.1038/nature08390.
- Pasman, Z, M D Been, and M A Garcia-Blanco. 1996. "Exon Circularization in Mammalian Nuclear Extracts." *RNA* 2 (6): 603–10.
- Patro, Rob, Stephen M. Mount, and Carl Kingsford. 2014. "Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms." *Nature Biotechnology* 32 (5): 462–64. doi:10.1038/nbt.2862.
- Pelletier, Jerry, and Nahum Sonenberg. 1988. "Internal Initiation of Translation of Eukaryotic mRNA Directed by a Sequence Derived from Poliovirus RNA." *Nature* 334 (6180): 320–25. doi:10.1038/334320a0.
- Petavy, G., J. R. David, P. Gibert, and B. Moreteau. 2001. "Viability and Rate of Development at Different Temperatures in *Drosophila*: A Comparison of Constant and Alternating Thermal Regimes." *Journal of Thermal Biology* 26 (1): 29–39. doi:10.1016/S0306-4565(00)00022-X.
- Raney, Brian J., Melissa S. Cline, Kate R. Rosenbloom, Timothy R. Dreszer, Katrina Learned, Galt P. Barber, Laurence R. Meyer, et al. 2011. "ENCODE Whole-Genome Data in the UCSC Genome Browser (2011 Update)." *Nucleic Acids Research* 39 (Database issue): D871–75. doi:10.1093/nar/gkq1017.
- Roberts, Adam, and Lior Pachter. 2013. "Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments." *Nature Methods* 10 (1): 71–73. doi:10.1038/nmeth.2251.
- Rowley, Janet D. 1973. "A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia Identified by Quinacrine Fluorescence and Giemsa Staining." <http://www.nature.com/nature/journal/v243/n5405/abs/243290a0.html>.
- Sajdel-Sulkowska, Elizabeth M., Kosal Nguon, Zachary L. Sulkowski, and Boguslaw Lipinski. 2007. "Potential Role of Oxidative Stress in Mediating the Effect of Altered Gravity on the Developing Rat Cerebellum." *Advances in Space Research* 40 (9): 1414–20. doi:10.1016/j.asr.2007.08.004.
- Salzman, Julia, Raymond E. Chen, Mari N. Olsen, Peter L. Wang, and Patrick O. Brown. 2013. "Cell-Type Specific Features of Circular RNA Expression." *PLoS Genet* 9 (9): e1003777. doi:10.1371/journal.pgen.1003777.
- Salzman, Julia, Charles Gawad, Peter Lincoln Wang, Norman Lacayo, and Patrick O. Brown. 2012. "Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types." *PLOS ONE* 7 (2): e30733. doi:10.1371/journal.pone.0030733.

- Santos, Manuel A. S., and Mick F. Tuite. 1995. "The CUG Codon Is Decoded in Vivo as Serine and Not Leucine in *Candida Albicans*." *Nucleic Acids Research* 23 (9): 1481–86. doi:10.1093/nar/23.9.1481.
- Sonnenfeld, Gerald, and William T Shearer. 2002. "Immune Function during Space Flight." *Nutrition* 18 (10): 899–903. doi:10.1016/S0899-9007(02)00903-6.
- Sooknanan, Roy, Jim Pease, and Ken Doyle. 2010. "Novel Methods for rRNA Removal and Directional, Ligation-Free RNA-Seq Library Preparation." *Nature Methods* 7 (10).
<http://www.nature.com/nmeth/journal/v7/n10/full/nmeth.f.313.html>.
- Stein, T. P. 2002. "Space Flight and Oxidative Stress." *Nutrition* 18 (10): 867–71. doi:10.1016/S0899-9007(02)00938-3.
- Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. "The Cancer Genome." *Nature* 458 (7239): 719–24. doi:10.1038/nature07943.
- Suzuki, Hitoshi, and Toshifumi Tsukahara. 2014. "A View of Pre-mRNA Splicing from RNase R Resistant RNAs." *International Journal of Molecular Sciences* 15 (6): 9331–42. doi:10.3390/ijms15069331.
- Taylor, Katherine, Kurt Kleinhesselink, Michael D. George, Rachel Morgan, Tangi Smallwood, Ann S. Hammonds, Patrick M. Fuller, et al. 2014. "Toll Mediated Infection Response Is Altered by Gravity and Spaceflight in *Drosophila*." *PLOS ONE* 9 (1): e86485. doi:10.1371/journal.pone.0086485.
- Tembe, Waibhav D., Stephanie JK Pond, Christophe Legendre, Han-Yu Chuang, Winnie S. Liang, Nancy E. Kim, Valerie Montel, et al. 2014. "Open-Access Synthetic Spike-in mRNA-Seq Data for Cancer Gene Fusions." *BMC Genomics* 15: 824. doi:10.1186/1471-2164-15-824.
- Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. 2012. "Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks." *Nature Protocols*, comp protocol, 7 (3): 562–78. doi:10.1038/nprot.2012.016.
- Tyler, Mary. 2000. "Chapter 8: Development of the Fruit Fly, *Drosophila Melanogaster*." In *Developmental Biology, A Guide for Experimental Study*, 2nd ed. Sinhaer Associates.
<http://www.sdbonline.org/sites/archive/SDBEduca/TylerContPref.html>.
- Udvardy, A. 1993. "Purification and Characterization of a Multiprotein Component of the *Drosophila* 26 S (1500 kDa) Proteolytic Complex." *Journal of Biological Chemistry* 268 (12): 9055–62.
- Vernos, I., J. Gonzalez-Jurado, M. Calleja, and R. Marco. 2002. "Microgravity Effects on the Oogenesis and Development of Embryos of *Drosophila Melanogaster* Laid in the Spaceshuttle during the Biorack Experiment (ESA)." *International Journal of Developmental Biology* 33 (2): 213–26. doi:10.1387/ijdb.2518159.
- Versari, Silvia, Giulia Longinotti, Livia Barenghi, Jeanette Anne Marie Maier, and Silvia Bradamante. 2013. "The Challenging Environment on Board the International Space Station Affects Endothelial Cell Function by Triggering Oxidative Stress through Thioredoxin Interacting Protein Overexpression:

- The ESA-SPHINX Experiment.” *The FASEB Journal* 27 (11): 4466–75.
doi:10.1096/fj.13-229195.
- Walter, Nils G, and David R Engelke. 2002. “Ribozymes: Catalytic RNAs That Cut Things, Make Things, and Do Odd and Useful Jobs.” *Biologist (London, England)* 49 (5): 199–203.
- Warde-Farley, David, Sylva L. Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, et al. 2010. “The GeneMANIA Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function.” *Nucleic Acids Research* 38 (suppl 2): W214–20.
doi:10.1093/nar/gkq537.
- Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna M. Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. “The Cancer Genome Atlas Pan-Cancer Analysis Project.” *Nature Genetics* 45 (10): 1113–20. doi:10.1038/ng.2764.
- Yamao, F., A. Muto, Y. Kawauchi, M. Iwami, S. Iwagami, Y. Azumi, and S. Osawa. 1985. “UGA Is Read as Tryptophan in Mycoplasma Capricolum.” *Proceedings of the National Academy of Sciences of the United States of America* 82 (8): 2306–9.
- Yunis, J. J., and A. L. Soreng. 1984. “Constitutive Fragile Sites and Cancer.” *Science* 226 (4679): 1199–1204. doi:10.1126/science.6239375.

Appendix

Supplemental Figures and Tables for Chapter 2

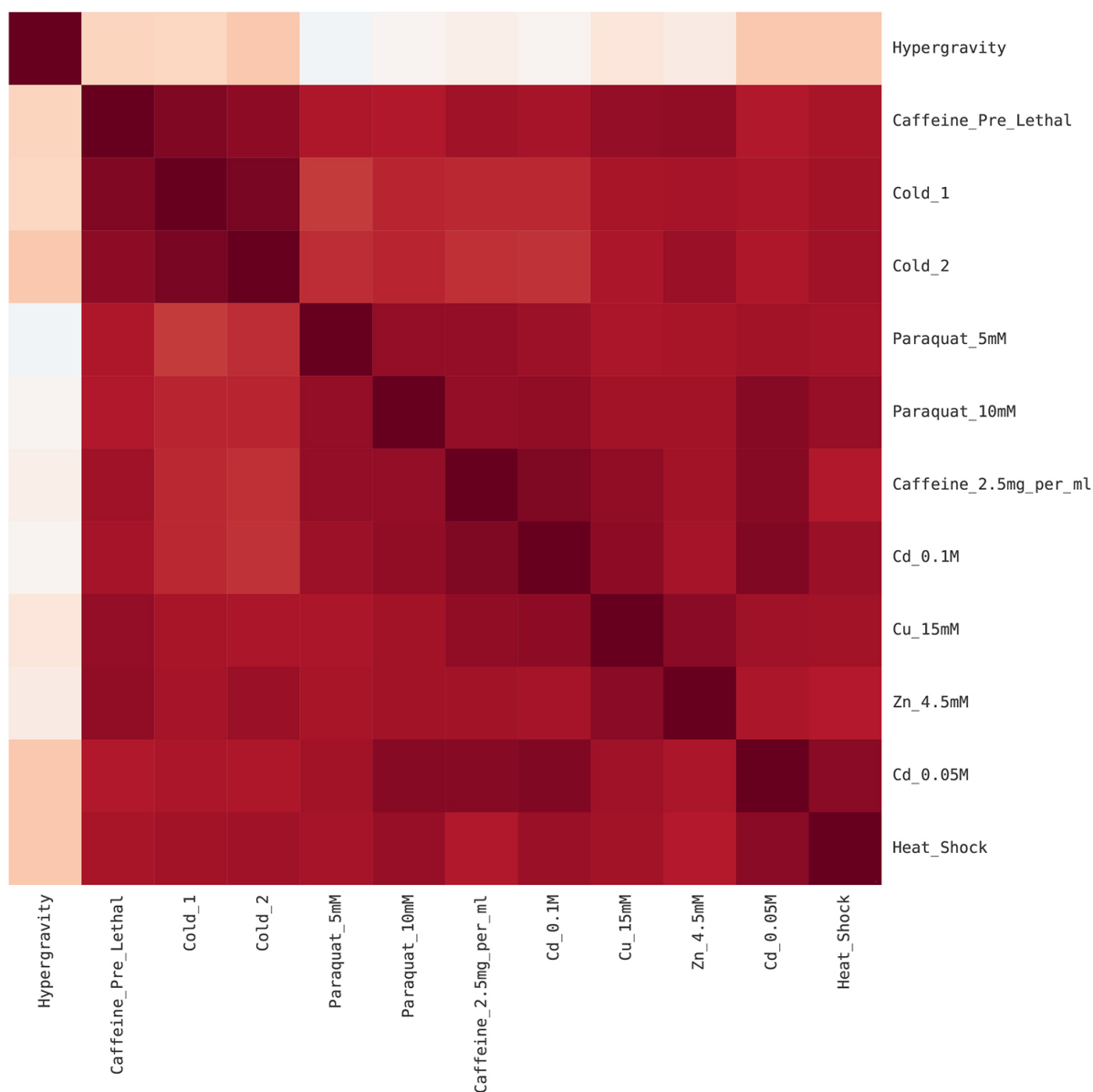


Figure A.1 Cluster map of hypergravity genes versus stress response genes

Hypergravity log2 fold change correlation with stress response fold change to stressors indicated. Darker red corresponds to greater correlation.

Table A.1 Positive PANTHER overrepresentation test for biological processes

Blue = ion channels, orange = redox, green = cuticle. PANTHER test version: release 20160715. Annotation: GO Ontology database 2016-06-22. Bonferroni correction = True.

GO biological process complete	control reference (12753)	DE genes (1407)	expected	over / under	fold enrich- ment	P-val
mitochondrial translation (GO:0032543)	21	14	2.32	+	6.04	4.37E-04
mitochondrion organization (GO:0007005)	138	42	15.23	+	2.76	2.57E-05
cellular component organization (GO:0016043)	2350	195	259.27	-	0.75	7.68E-03
cellular component organization or biogenesis (GO:0071840)	2400	198	264.78	-	0.75	3.86E-03
translation (GO:0006412)	279	68	30.78	+	2.21	7.95E-06
cellular macromolecule metabolic process (GO:0044260)	2166	143	238.97	-	0.6	7.25E-10
peptide biosynthetic process (GO:0043043)	285	68	31.44	+	2.16	1.80E-05
peptide metabolic process (GO:0006518)	336	71	37.07	+	1.92	8.35E-04
organonitrogen compound metabolic process (GO:1901564)	845	166	93.23	+	1.78	2.42E-09
cellular amide metabolic process (GO:0043603)	348	75	38.39	+	1.95	1.83E-04
organonitrogen compound biosynthetic process (GO:1901566)	477	100	52.63	+	1.9	4.70E-06
amide biosynthetic process (GO:0043604)	289	68	31.88	+	2.13	3.06E-05
potassium ion transport (GO:0006813)	38	16	4.19	+	3.82	2.21E-02
metal ion transport (GO:0030001)	141	37	15.56	+	2.38	6.09E-03
cation transport (GO:0006812)	140	36	15.45	+	2.33	1.28E-02
ion transport (GO:0006811)	220	55	24.27	+	2.27	1.14E-04
monovalent inorganic cation transport (GO:0015672)	315	77	34.75	+	2.22	6.26E-07
transmembrane transport (GO:0055085)	389	73	42.92	+	1.7	3.46E-02
cellular respiration (GO:0045333)	73	26	8.05	+	3.23	9.58E-04
energy derivation by oxidation of organic compounds (GO:0015980)	84	29	9.27	+	3.13	3.94E-04
generation of precursor metabolites and energy (GO:0006091)	116	34	12.8	+	2.66	1.46E-03
oxidation-reduction process (GO:0055114)	478	92	52.74	+	1.74	9.28E-04
single-organism metabolic process (GO:0044710)	1289	195	142.21	+	1.37	1.33E-02
ATP metabolic process (GO:0046034)	80	24	8.83	+	2.72	4.45E-02
purine ribonucleoside triphosphate metabolic process (GO:0009205)	84	25	9.27	+	2.7	3.42E-02
ribonucleoside triphosphate metabolic process (GO:0009199)	85	25	9.38	+	2.67	4.15E-02
nucleobase-containing compound metabolic process (GO:0006139)	1114	64	122.9	-	0.52	2.53E-06

organic cyclic compound metabolic process (GO:1901360)	1264	87	139.45	-	0.62	1.09E-03
heterocycle metabolic process (GO:0046483)	1177	74	129.85	-	0.57	5.37E-05
cellular aromatic compound metabolic process (GO:0006725)	1221	78	134.71	-	0.58	5.96E-05
purine nucleoside triphosphate metabolic process (GO:0009144)	84	25	9.27	+	2.7	3.42E-02
chitin-based cuticle development (GO:0040003)	187	45	20.63	+	2.18	5.03E-03
single-organism developmental process (GO:0044767)	3041	266	335.5	-	0.79	1.29E-02
Unclassified (UNCLASSIFIED)	2536	302	279.79	+	1.08	0.00E+0

Table A.2 Positive PANTHER overrepresentation test for molecular function

Blue = ion channels, orange = redox, green = cuticle, red = other. PANTHER test version: release 20160715.
 Annotation: GO Ontology database 2016-06-22. Bonferroni correction = True.

GO molecular function complete	control reference (12753)	DE genes (1407)	expected	over / under	fold enrich- ment	P-val
voltage-gated potassium channel activity (GO:0005249)	21	12	2.32	+	5.18	6.56E-03
voltage-gated cation channel activity (GO:0022843)	29	17	3.2	+	5.31	5.70E-05
cation channel activity (GO:0005261)	111	34	12.25	+	2.78	2.34E-04
cation transmembrane transporter activity (GO:0008324)	322	78	35.53	+	2.2	3.04E-07
ion transmembrane transporter activity (GO:0015075)	456	97	50.31	+	1.93	1.73E-06
substrate-specific transmembrane transporter activity (GO:0022891)	576	113	63.55	+	1.78	6.38E-06
substrate-specific transporter activity (GO:0022892)	632	119	69.73	+	1.71	2.29E-05
transporter activity (GO:0005215)	782	137	86.28	+	1.59	1.22E-04
transmembrane transporter activity (GO:0022857)	670	122	73.92	+	1.65	8.97E-05
ion channel activity (GO:0005216)	210	48	23.17	+	2.07	3.88E-03
substrate-specific channel activity (GO:0022838)	215	49	23.72	+	2.07	3.33E-03
channel activity (GO:0015267)	224	49	24.71	+	1.98	9.70E-03
passive transmembrane transporter activity (GO:0022803)	224	49	24.71	+	1.98	9.70E-03
voltage-gated ion channel activity (GO:0005244)	38	20	4.19	+	4.77	2.26E-05
voltage-gated channel activity (GO:0022832)	38	20	4.19	+	4.77	2.26E-05
gated channel activity (GO:0022836)	155	42	17.1	+	2.46	2.43E-04
potassium channel activity (GO:0005267)	27	13	2.98	+	4.36	1.64E-02
potassium ion transmembrane transporter activity (GO:0015079)	43	18	4.74	+	3.79	2.85E-03
metal ion transmembrane transporter activity (GO:0046873)	188	48	20.74	+	2.31	1.90E-04
inorganic cation transmembrane transporter activity (GO:0022890)	257	66	28.35	+	2.33	7.74E-07
monovalent inorganic cation transmembrane transporter activity (GO:0015077)	200	52	22.07	+	2.36	3.37E-05
structural constituent of ribosome (GO:0003735)	172	61	18.98	+	3.21	9.12E-12
structural molecule activity (GO:0005198)	492	123	54.28	+	2.27	1.73E-13

structural constituent of chitin-based larval cuticle (GO:0008010)	104	35	11.47	+	3.05	1.68E-05
structural constituent of chitin-based cuticle (GO:0005214)	136	41	15	+	2.73	2.10E-05
structural constituent of cuticle (GO:0042302)	141	43	15.56	+	2.76	6.90E-06
oxidoreductase activity (GO:0016491)	534	103	58.91	+	1.75	6.97E-05
Unclassified (UNCLASSIFIED)	2690	328	296.78	+	1.11	0.00E+0

Table A.3 PANTHER enrichment test for biological processes

PANTHER test version: release 20160715. Annotation: GO Ontology database 2016-06-22. Bonferroni correction = True.

GO biological process complete	number	over / under	P-val
chromatin assembly or disassembly (GO:0006333)	25	-	3.84E-08
chromatin organization (GO:0006325)	32	-	1.14E-06
chromosome organization (GO:0051276)	33	-	2.84E-06
organelle organization (GO:0006996)	127	-	1.89E-05
cellular component organization (GO:0016043)	195	-	4.02E-03
cellular component organization or biogenesis (GO:0071840)	198	-	3.70E-03
macromolecular complex subunit organization (GO:0043933)	59	-	1.31E-07
response to bacterium (GO:0009617)	36	+	2.27E-03
defense response to other organism (GO:0098542)	36	+	5.97E-03
defense response (GO:0006952)	51	+	2.36E-04
response to other organism (GO:0051707)	50	+	6.49E-05
response to external biotic stimulus (GO:0043207)	50	+	6.49E-05
response to biotic stimulus (GO:0009607)	50	+	6.49E-05
cellular respiration (GO:0045333)	26	-	6.59E-03
energy derivation by oxidation of organic compounds (GO:0015980)	29	-	3.05E-02
cellular process (GO:0009987)	629	-	4.23E-02
myofibril assembly (GO:0030239)	10	-	9.97E-03
actomyosin structure organization (GO:0031032)	13	-	2.46E-02
cellular component assembly involved in morphogenesis (GO:0010927)	13	-	2.74E-03
metamorphosis (GO:0007552)	46	+	1.12E-02
proteolysis (GO:0006508)	77	+	4.40E-02
defense			
redox			
proteolysis			

Table A.4 PANTHER enrichment test for molecular function

PANTHER test version: release 20160715. Annotation: GO Ontology database 2016-06-22. Bonferroni correction = True.

GO molecular function complete	number	over /under	P-val
serine-type endopeptidase activity (GO:0004252)	35	+	3.08E-08
endopeptidase activity (GO:0004175)	47	+	7.73E-07
peptidase activity, acting on L-amino acid peptides (GO:0070011)	72	+	1.15E-03
peptidase activity (GO:0008233)	72	+	1.15E-03
serine-type peptidase activity (GO:0008236)	39	+	6.11E-06
serine hydrolase activity (GO:0017171)	39	+	6.11E-06
protein heterodimerization activity (GO:0046982)	32	-	2.91E-04
protein dimerization activity (GO:0046983)	43	-	1.07E-02
protein binding (GO:0005515)	144	-	9.91E-03
proteolysis			

Table A.5 Gene expression data (qRT-PCR) at 3 g, 5 g and 8 g

Red font indicates significant change, p-value < 0.05.

Genes	mRNA Seq (Log2 fold change of 3 g over 1 g)	qRT-PCR (Log2 fold change over 1g)		
		3 g	5 g	8 g
Cpr65Ec	3.27	8.32	-0.18	0.28
Cpr97Ea	-1.405	-1.51	1.56	0.23
Cpr97Eb	-1.929	-2.64	1.51	0.49
Cpr62Bc	-4.003	-5.05	1.66	0.67
Cpr64Ab	4.786	6.93	1.65	7.39
Cpr64Ad	3.031	5.10	6.23	5.74
Cda5	-1.359	-1.39	1.47	0.44
Cht4	3.308	4.26	-0.55	-0.168
Edg78E	3.084	6.43	5.84	-0.57
CG32679	3.031	9.87	0.67	1.04
Drsl6	1.741	6.92	-0.014	-0.62
Drs11	2.528	7.76	1.48	-0.83
CG16995	1.696	2.20	0.042	1.54
Eig71Eh	6.843	10.97	-1.18	-0.57
Eig71Ec	8.37	11.31	-1.25	-0.53
Eig71Eg	7.129	11.01	-0.76	0.20
Eig71Ea	8.053	10.18	-0.73	-0.35
TotA	1.864	3.00	2.98	1.95

Table A.6 Software details

Step	Program_and_version	Settings	notes
remove reads not passing Illumina Casava filter	casava illumina filter 0.1 05-Aug-2011	fastq_illumina_filter --keep N -v -v -o ../read1_filtered.fastq read1.fastq	
check read quality	fastqc	fastqc filename	
trim adapters and low quality reads	Trimmomatic-0.30	trimmomatic.jar PE -phred33 read1_filtered.fastq read2_filtered.fastq read1_trimmed_paired.fastq.gz read1_trimmed_unpaired.fastq.gz read2_trimmed_paired.fastq.gz read2_trimmed_unpaired.fastq.gz ILLUMINACLIP:../TruSeq3-PE.fa:2:30:10:8:TRUE TRAILING:20 MINLEN:35	adapters: >PrefixPE/1 TACACTCTTTCCCTACACGACGCTCTTCCGATCT, >PrefixPE/2 GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT, >PE1 TACACTCTTTCCCTACACGACGCTCTTCCGATCT, >PE1_rc AGATCGGAAGAGCGTCGTGTAGGAAAGAGTGTAA, >PE2 GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT, >PE2_rc AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC, >postindex ATCTCGTATGCCGTCTTCTGCTTG, >postindex_rc CAAGCAGAAGACGGCATACGAGAT
build reference for bowtie2	bowtie2 version2.1.0	bowtie2-build --offrate 1 dmel-all-transcript-r6.02.fasta dmel-all-transcript-r6.02	reference from ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.02_FB2014_05/fasta/dmel-all-transcript-r6.02.fasta.gz
map reads	bowtie2 version2.1.0	bowtie2 -p 16 -t -q -a --un-gz unmapped.fastq.gz -x dmel-all-transcript-r6.02 -1 read1_trimmed_paired.fastq.gz -2 read2_trimmed_paired.fastq.gz -U read1_trimmed_unpaired.fastq.gz,read2_trimmed_unpaired.fastq.gz samtools view -Sb - > bt2out.bam	
calculate abundance	eXpress 1.5.1	express dmel-all-transcript-r6.02.fasta bt2out.bam	
sum transcripts to gene level using TPM * 100			
DE analysis	DESeq2	followed the DESeq2 vignette with alpha = 0.05	
GO analysis	Panther 9.0 release		

Table A.7 qRT-PCR primers

Gene	Forward_Primer	Reverse_Primer
Cpr65Ec	F: 5'- CGCGAGTATAAGAGCGACCT- 3'	R: 5' -CGCTGGCATAGTAACCACCC- 3'
Cpr97Ea	F: 5'- CAACAGCACCACCAACCATC-3'	R: 5'- CTGCGTTTGCGAGTAATCCG-3'
Cpr97Eb	F: 5'- GAAATTGACCTTGAGCTTGGGT -3'	R: 5'- CGTTGTGCTTGTCGATCTGC-3'
Cpr62Bc	F: 5'- ATGACCACAATGGCTTCAACG-3'	R: 5'- AGGGATCCGGCGTAGGG-3'
Cpr64Ab	F: 5'- CGCGTTCGCCTATAATGTGC-3'	R: 5'- GAGCCCTTGACCACATCTCC-3'
Cpr64Ad	F: 5'- GTGCTCCTCTGGCTTACTCC-3'	R: 5'- CGTAGGTCTTGGACACCAGG-3'
Cda5	F: 5'- TGCTGGGATTGTGCCTACTG-3'	R: 5'- TGTCGGTCTTGATGTTGGTCC-3'
Cht4	F: 5'- CCACTTCCCTGGCCTTCATT-3'	R: 5'- AATCTCGCGCAGCAGAGTAA-3'
Edg78E	F: 5'- CAGTACGCCTACGAGACCAG-3'	R: 5'- GGAGAACGTAAGCCGGAAC-3'
CG32679	F: 5'- TGCGATCCAGAACTGTGTCC-3'	R: 5'- GGGAAAGCCACTAACTCCACC-3'
PGRP-SB2	F: 5'- CCCTTGCTTCAATCCGCATC-3'	R: 5'- GCCCTCGTAGATTTCGACCAT-3'
Drsl6	F: 5'- CTCGCTCTGCTGATGATGGTC-3'	R: 5'- TGCATGTCTCGTTGTCCCA-3'
Drsl1	F: 5'- TCCTGTCCGCTGTCTTGATG-3'	R: 5'- TACGCACTTCTTCTGTGCC-3'
AttD	F: 5'- AGCCCAAGCCAATCTCTTCC-3'	R: 5'- TCATGCCGAAGTGAAGAAC-3'
Scpr-A	F: 5'- GCGTTCACCAAGGTTCTCCA-3'	R: 5'- AGTTTGTGATGCGGCTCGAT-3'
CG16995	F: 5'- AAGACAAGTGCCAGCGAAAG-3'	R: 5'- TCAAGGGTTGAGCACCATGC-3'
Eig71Eh	F: 5'- TGTCTGCTTCCTGGTGATCC-3'	R: 5'- CGCAAAGATTCTGGTCCCTTCC-3'
Eig71Ec	F: 5'- ACTCTCATTTTGTCTATTCTCTGCC-3'	R: 5'- TCCGTTCAATTCGACGGCAA-3'
Eig71Eg	F: 5'- TGGTATTTTGGCTTCTGCTGC-3'	R: 5'- GGTTGGCTGTTACGAAGAC-3'
Eig71Ea	F: 5'- AGAGTGCAGACAAAGGACCG-3'	R: 5'- AGTTCATGGGTGATTGCCAAC-3'
TotA	F: 5'- TCAACTGCTCTTATGTGCTTTGC-3'	R: 5'- CTCACGATCTTCGTCGGAATAG-3'
CG44625	F: 5'- GAATCGTGCTGTTCGCCATT-3'	R: 5'- GTGGAGGCTGTGGTATTCCG-3'
Tubulin	F: 5'- TGTCGCGTGTGAAACACTTC-3';	R: 5'-AGCAGGCGTTTCCAATCTG-3'