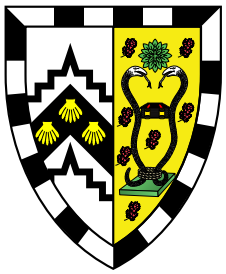
Quantitative genetics of gene expression during fruit fly development

Nils Kölling

European Bioinformatics Institute
Gonville and Caius College
University of Cambridge



This dissertation is submitted for the degree of

Doctor of Philosophy

August 2015

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the Degree Committee for the Faculty of Biology.

Abstract

Over the last ten years, genome-wide association studies (GWAS) have been used to identify genetic variants associated with many diseases as well as quantitative phenotypes, by exploiting naturally occurring genetic variation in large cohorts of individuals. More recently, the GWAS approach has also been applied to high-throughput RNA sequencing (RNA-seq) data in order to find loci associated with different levels of gene expression, called expression quantitative trait loci (eQTL).

Because of the large amount of data that is required for such high-resolution eQTL studies, most of them have so far been carried out in humans, where the cost of data collection could be justified by a possible future impact in human health. However, due to the rapidly falling price of high-throughput sequencing it is now also becoming feasible to perform high-resolution eQTL studies in higher model organisms. This enables the study of gene regulation in biological contexts that have so far been beyond our reach for practical or ethical reasons, such as early embryonic development.

Taking advantage of these new possibilities, we performed a high-resolution eQTL study on 80 inbred fruit fly lines from the *Drosophila* Genetic Reference Panel, which represent naturally occurring genetic variation in a wild population of *Drosophila melanogaster*. Using a $3'$ Tag RNA-sequencing protocol we were able to estimate the level of expression both of genes as well as of different $3'$ isoforms of the same gene. We estimated these expression levels for each line at three different stages of embryonic development, allowing us to not only improve our understanding of *D. melanogaster* gene regulation in general, but also investigate how gene regulation changes during development.

In this thesis, I describe the processing of $3'$ Tag-Seq data into both $3'$ isoform expression levels and overall gene expression levels. Using these expression levels I call proximal eQTLs both common and specific to a single developmental stage with a multivariate linear mixed model approach while accounting for various confounding factors. I then investigate the properties of these eQTLs, such as their location or the gene categories enriched or depleted in eQTLs. Finally, I extend the proximal eQTL calling approach to distal variants to find gene regulatory mechanisms acting in *trans*.

Taken together, this thesis describes the design, challenges and results of performing a multivariate eQTL study in a higher model organism and provides new insights into gene regulation in *D. melanogaster* during embryonic development.

Acknowledgments

I am extremely thankful to all the people who have made my PhD such a great experience.

First, of course, a big thank you to Ewan Birney, who gave me the chance to work in his research group and supported me throughout my PhD and beyond. This project would not have been possible without his expertise, guidance and limitless enthusiasm. I would also like to thank Ian Dunham, who co-supervised me at the beginning of my PhD and helped me get this project off the ground. My thanks also to all the other members of the Birney research group, with whom I had many productive discussions about this project: Sander Timmer, Mikhail Spivakov, Sandro Morganella, Valentina Iotchkova, Helena Kilpinen, Hannah Meyer, Leland Taylor and Dirk Dolle. It was great sharing an office with you! Thank you also to Stacy Knoop and the rest of the A-team for organising my travels and helping me find time with Ewan despite his busy schedule.

My PhD project was based on a very productive collaboration with Eileen Furlong's group at EMBL Heidelberg and in particular the work of Enrico Cannavò. I would like to thank Enrico for the many days and nights he spent in the lab collecting data for this project, as well as him, Eileen and the other members of her group for the many fruit(fly)ful discussions that we had over the years. In addition, many thanks to Paolo Casale and Oliver Stegle from the EBI for all their advice on the statistical analysis and their help with setting up LIMIX for this project. My thesis advisory committee — John Marioni, Jeff Barrett and Jan Korbel — were also extremely helpful with their suggestions and constructive criticism. Thank you for your advice and keeping Ewan in check!

While writing this thesis, I was supported by a team of proofreaders who were reading chapters faster than I could write them. A big thank you for all the feedback, suggestions and your (usually) encouraging remarks, Maria Xenophontos, Myrto Kostadima (sorry about the Zs…), Steve Wilder (… and the Us), Ângela Gonçalves (fortunately!), Konrad Rudolph (also for all the "typesetting" help), Hannah Meyer, Valentina Iotchkova, Helena Kilpinen and Paolo Casale!

What made EMBL-EBI such an amazing place to do my PhD was not just the science, but also the great community of PhD students, alumni and "honorary" PhD students. I would like to thank all of them for the long breakfasts in the DiNA, the many lunches in Murray's and everything else! You all played a big role in keeping me (relatively) sane during my PhD, in particular during the thesis writing. I would especially like to thank Maria and Konrad for being great housemates and enduring my jokes of sometimes questionable quality, day after day. Whether we were having scientific discussions in the middle of the night, organised parties or spent the day coding in the kitchen, I always enjoyed my time here. I will miss the Mansion! I would also like to thank my fellow EBI PhD students from the class of 2015(ish), Konrad, Tom, Michael, Kevin and Ewan, as well as all the other EMBL PhD students, particularly Ola, Joana, Katya and Thibaut. Thank you for the great company during the predoc course (including countless foosball games) and being excellent hosts every time I came back to Heidelberg!

Finally, I would like to thank my brother Jannes and my parents Heike and Ralf, who got me excited about technology and science from an early age and always supported me during my studies. Vielen, vielen Dank!

Contents

List of common abbreviations

3′ Tag-Seq	3′ Tag Sequencing
3′i-eQTL	3′ isoform eQTL
A	Adenine
APA	Alternative Polyadenylation
apaQTL	Alternative Polyadenylation QTL
BH	Benjamini & Hochberg
bp	base pairs
C	Cytosine
cDNA	complementary DNA
CRM	*Cis* Regulatory Module
DGRP	*Drosophila* Genetic Reference Panel
DHS	DNase I Hypersensitive Site
DNA	Deoxyribonucleic Acid
DSE	Downstream Sequence Element
eQTL	expression QTL
eQTN	expression QTN
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase of transcript per Million reads mapped
G	Guanine
GO	Gene Ontology
GWAS	Genome-Wide Association Study
Indel	Insertion/deletion
is-eQTL	isoform-specific eQTL
kb	kilo base pairs (1,000 bp)
LD	Linkage Disequilibrium
Mb	mega base pairs (1,000 kb)
miRNA	micro RNA
mRNA	messenger RNA
MZT	Maternal to Zygotic Transition

ncRNA	non-coding RNA
PCA	Principal Component Analysis
pre-mRNA	precursor mRNA
QTL	Quantitative Trait Locus
QTN	Quantitative Trait Nucleotide
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
SNP	Single-Nucleotide Polymorphism
SV	Structural Variation
T	Thymine
TPM	Transcripts Per Million
TSS	Transcription Start Site
U	Uracil
USE	Upstream Sequence Element
UTR	Untranslated Region

1. Introduction

In 1865, Gregor Mendel laid the foundations for the systematic study of inheritance with his famous experiments on pea plants. In the 150 years since then, techniques such as linkage mapping and genome-wide association studies have identified genetic variation associated with thousands of different traits and diseases. Yet, despite this extensive amount of research, the molecular mechanisms through which differences between genomes result in differences between whole organisms remain poorly understood.

To bridge this gap between genotype and phenotype, we need to understand the consequences of genetic variation at the cellular level. A major factor in this is gene regulation, in which the expression level of genes is adjusted according to regulatory signals encoded in the genome. Over the last decade, high-resolution expression quantitative trait locus (eQTL) studies, based on new methods for quantifying gene expression, have emerged as a major avenue for studying this process.

To date, most such eQTL studies have been conducted in humans, in particular in the context of human health. However, the same concepts are also applicable to model organisms kept in the laboratory, where breeding patterns and environmental conditions can be tightly controlled. This enables the study of gene regulation not only in different environmental conditions, but also in different stages of an animal's life.

A crucial stage of life is embryonic development, when the body plan of the organism is laid out and cells begin to form different tissues. Genetic differences that affect these fundamental processes can have major consequences on the adult organism, even if the processes themselves are no longer active later in life. Thus, by studying model organisms during early development, we can uncover important mechanisms that lie beyond the reach of studies in humans.

In this dissertation, I will describe an eQTL study across multiple stages of embryonic development in the model organism *Drosophila melanogaster*. I will begin this introduction by giving an overview of the history of genetics with a

particular focus on the role played by *Drosophila* in many important discoveries. This will be followed by a closer description of the topics relevant to this project, including *Drosophila* embryo development, gene regulation, the estimation of gene expression levels, genetic association studies and the *Drosophila* Genetic Reference Panel. Finally, I will describe, in detail, the experimental design of this project.

1.1. From yellow peas to quantitative genetics

1.1.1. The principles of inheritance

In *On the Origin of Species* (Darwin, 1859), Charles Darwin proposed the theory of evolution, which rests on the principles that natural variation between individuals provides differential reproductive advantages, and that this variation is heritable. While his theory could beautifully explain the adaptation of a population to its environment and the development of new species, the mechanisms by which such variation might occur and how it could be passed on from generation to generation was not clear. In the words of Darwin in *On the Origin of Species*: "Our ignorance of the laws of variation is profound" and "[t]he laws governing inheritance are quite unknown".

Unbeknownst to Darwin, the Austrian friar Gregor Mendel had started working on exactly this problem in 1853. By carefully breeding pea plants with different traits (such as seed colour and shape) in a controlled environment, Mendel was able to study how these traits were passed on from parent generations to their offspring. For example, he bred plants with yellow and with green seeds and then observed the seed colour of their offspring. In the first (F_1) generation, all of the seeds were yellow. However, when he bred the plants from the yellow F_1 generation with each other, he observed that approximately a quarter of the next generation (F_2) had green seeds, while the rest of the seeds were yellow. Thus, he discovered the principle of (and coined the terms for) dominant (yellow) and recessive (green) traits.

His research led Mendel to propose his famous Laws of Inheritance: the Law of Segregation (which describes how each individual contains a pair of alleles for any given trait, one of which is passed on to its offspring at random) and the Law of Independent Assortment (which describes how different traits are inherited independently from each other).

Unfortunately, while Mendel published this work in his 1866 paper *Versuche über Pflanzenhybriden* (Mendel, 1866), it stayed largely unnoticed and Darwin

is said to have been unaware of it. It took several more decades until, in 1900, Hugo de Vries and Carl Correns independently rediscovered and popularised the principles Mendel had described. After this, the school of Mendelianism became increasingly popular and scientists started to work on identifying the molecular basis of Mendel's laws.

In December 1901, William Bateson, who had learned of Mendel through de Vries's publications, introduced Mendel's laws at the Royal Society's Evolution Committee. It was in this lecture that he introduced some fundamental terms of genetics that are still in use today, such as "allelomorph" (allele), "zygote", "homozygous", "heterozygous" and, indeed, the word "genetics" itself.

An important step towards reconciling Mendel's laws of inheritance with Darwin's theory of evolution was made in 1902, when Theodor Boveri showed, in sea urchin, that different chromosomes contained different hereditary material and an organism required a full haploid set of them to function. In 1903, Walter Sutton published a paper proposing how these principles, together with the random segregation of paternal and maternal chromosomes during gamete formation could form the molecular basis for Mendel's Laws of Inheritance (Sutton, 1903). Importantly, he also noted how the number of traits was much larger than the number of chromosomes, which meant that some traits had to be located on the same chromosome and be transmitted together.

1.1.2. *Drosophila* and the birth of modern genetics

These discoveries sparked a whole new era of biology, with attempts being made to find both the molecular mechanisms that could bring about genetic variation (mutations) as well as the mechanisms that could lead to their inheritance.

Due to its quick generation time of 12 days, the ease with which it could be bred and the simplicity of identifying differences in traits, the fruit fly *Drosophila* became an organism of choice for the study of mutations. The first experiments with this model organism were reported in 1906 (Castle et al., 1906) but the most important studies of *Drosophila* in the early 20^{th} century were conducted in the famous fly room of Thomas Hunt Morgan.

In 1909, Morgan was attempting to induce mutations in flies using different temperature ranges, as well as X-rays and radium (Sturtevant, 1959, page 293). He was sceptical of both Darwin's theory of natural selection and Sutton's proposal of chromosomes for the transmission of heredity, and was particularly critical of the suggestion that chromosomes could be involved in sex determination. This

changed, however, in 1910, when he discovered a single male fly with white eyes instead of the normal red eye colour. Initially, he did not think much of this mutation as mating this fly with red-eyed females resulted in very few white-eyed flies in the F_1 generation (only 3 out of 1,240 offspring, which he attributed to further random mutations). However, breeding the white-eyed male with females from the F_1 generation resulted in an F_2 generation with approximately 25 % white-eyed and 75 % red-eyed flies, as Mendel's laws would have predicted for a dominant red and a recessive white eye colour trait (Morgan, 1910). The discovery of this Mendelian trait itself was already interesting, but the most important discovery was that it occurred exclusively in males.

In his 1911 paper (Morgan, 1911a) Morgan proposed how this sex-limited inheritance could be explained if the factor leading to white eye colour was not only recessive but also attached, or *linked*, to the sex-determining factor on the X chromosome. This theory perfectly explained how all the female flies in the F_2 generation had to have red eye colour, as one of their X chromosomes must have come from a male F_1 fly, all of which carried the dominant red factor. At the same time, the male flies, only receiving one copy of the X chromosome from their mother, would randomly receive either the copy inherited from their red-eyed grandmother or the copy inherited from their white-eyed grandfather, resulting in half of them having red and half of them having white eyes. Today, the gene implicated in this mutation is still known as *white* (*w*) and is, of course, located on the X chromosome.

Together with his students Hermann Muller, Alfred Sturtevant and Calvin Bridges, Morgan went on to discover and investigate many more mutations in *Drosophila*, which helped to uncover several important principles of genetics still relevant today. Among these was the observation that the offspring of female flies with two X-linked mutations on separate chromosomes sometimes carried both mutations on a single X chromosome. This lead him to propose the concept of crossing over, the exchange of genetic material between homologous chromosomes during meiosis (Morgan, 1911b). Morgan also reasoned that the degree of coupling between two regions would be relative to their linear distance on the chromosome. He and his students used this phenomenon to develop the technique of gene mapping, using the recombination rate between different traits to estimate the relative distances of the genes from each other.

The first genetic map, which described the arrangement of genes on the X chromosome, was published in 1913 (Sturtevant, 1913). Two years later, Morgan,

Sturtevant, Muller and Bridges published a map covering chromosomes X, 2 and 3 as part of their text book entitled *The Mechanism of Mendelian Heridity* (Morgan et al., 1915). Bridges continued to focus on gene mapping in the following years (Bridges, 1916), developing standardised reagents to allow increasingly detailed mapping. Many of the genes and their alleles that these pioneers discovered are still actively under investigation today. For example, one of the genes Bridges discovered as a reference point for his work was *Dichaete* (Bridges and Morgan, 1923), which would become the first SOX domain protein to be identified in *Drosophila* (Russell et al., 1996).

Muller, together with Altenburg, another student of Morgan's, went on to use genetic linkage to show that a mutation leading to truncated wings was actually caused by multiple factors on different chromosomes, with the effect of one "master" mutation being modulated by additional mutations on different chromosomes, all of which were inherited in a Mendelian fashion (Altenburg and Muller, 1920). This discovery is an example of the concept of quantitative traits, which were formalised by R.A. Fisher in 1918 (see Section 1.1.3).

Due to the tremendous success Morgan had had with *Drosophila*, it quickly became the model organism of choice for many geneticists around the world. In 1933, Morgan received the Nobel Prize in Physiology or Medicine "for his discoveries concerning the role played by the chromosome in heredity".

1.1.3. Biometrics and population genetics

While Mendel was studying the inheritance of traits in peas in the 19$^{\text{th}}$ century, others were trying to quantify inheritance in the context of human traits.

One investigator among them was Francis Galton, a half-cousin of Darwin's, who started working on Darwin's theory of evolution and its implications shortly after its publication. He was particularly fascinated by the question of how evolution applied to humanity and how its effects could be used to improve the human race. To this end, Galton applied himself to the study of biometrics, trying to measure and estimate the heritability of human traits such as height and mental capabilities. Some of the concepts and methods he developed during these studies are still fundamental to genetics today (Galton, 1909). These include the concepts of correlation, regression toward the mean and the regression line, which Galton used to compare the heights of children to those of their parents. Galton's protégé was the mathematician Karl Pearson, who worked together with Galton to make several more important contributions to statistics. Among others, he introduced

the concepts of the p-value and the χ^2 test (Pearson, 1900) and proposed principal component analysis (PCA, Pearson, 1901).

In 1918, building on the work of Galton and Pearson, the statistician Ronald A. Fisher described how Mendelian inheritance could result in the continuous variation of a trait (Fisher, 1918). This work not only introduced the concepts of variance and analysis of variance (ANOVA) but also laid the foundation for the concept of quantitative traits and quantitative genetics. While working as a statistician at the Rothamsted Experimental Station, Fisher employed these concepts to study a large set of data that had been produced by the agricultural research institute over many decades. For example, he investigated the effects of different types of fertiliser on wheat yield, using a data set that covered the yield of 13 differently treated plots of land over more than 60 years. This work resulted in his series of publications entitled *Studies in Crop Variation* (see for example Fisher, 1921; Fisher and Mackenzie, 1923).

In 1930, Fisher published his seminal work *The Genetical Theory of Natural Selection*, which finally united the fields of Mendelian genetics and evolution through natural selection (Fisher, 1930). Thus, together with J.B.S. Haldane and Sewall Wright, Fisher essentially founded the field of population genetics in the 1930s.

1.2. *Drosophila melanogaster* as a model for development

Since the days of Morgan, *Drosophila* has remained an important model organism, particularly in the context of genetics and heritability. Initial studies were mostly concerned with mutations in adult flies until, in 1937, D. F. Poulson described how genetics (in the form of chromosomal deficiencies) affected the development of the *Drosophila* embryo (Poulson, 1937). This not only established the concept that early development in *Drosophila* was affected by genetics, but also established the framework in which these effects could be studied. Poulson and others continued to use this approach to study embryogenesis. However, *Drosophila* embryology remained a niche field for several more decades, largely due to the difficulties involved in working with the small embryos.

The study of *Drosophila* embryogenesis only truly became popular in the 1970s and 1980s, when technical improvements enabled the fixation of eggs for histological analysis without damaging them. This made it possible to study early embryogenesis in much greater detail than before (Turner and Mahowald, 1976).

These improvements enabled Christiane Nüsslein-Volhard and Eric Wieschaus,

while working at EMBL Heidelberg in 1980, to perform a genetic screen for mutations that changed the segmentation pattern of the *Drosophila* embryo (Nüsslein-Volhard and Wieschaus, 1980). In this historic screen, Nüsslein-Volhard and Wieschaus identified 15 mutations (including famous loci such as *even-skipped*, *engrailed* and *Krüppel*), which they associated with three different types of effects on segmentation patterns — segment polarity mutants (deletions of parts of each segment that are replaced by a mirror-image of the remainder), pair-rule mutants (deletions in alternating segments) and gap mutants (deletions of a whole stretch of segments). In 1995, Edward B. Lewis, Christiane Nüsslein-Volhard and Eric F. Wieschaus received the Nobel Prize in Physiology or Medicine for "their discoveries concerning the genetic control of early embryonic development".

Drosophila development has been extensively studied in the years since then and is now very well described. Under normal conditions, it takes approximately 22 h from fertilisation for the *Drosophila* embryo to develop into a larva. During this period, the cells of the embryo rapidly divide and differentiate, forming the precursors for the organs and appendages of the adult fly. The fly then develops for a further 4–5 days as a larva, followed by 5 days of metamorphosis as a pupa and finally the hatching of the adult fly (Weigmann et al., 2003).

1.2.1. The development of the *D. melanogaster* embryo

In the following sections, I will give an overview over the development of the *D. melanogaster* embryo in 2 h intervals, based on detailed descriptions available in Campos-Ortega and Hartenstein (1997), Brody (1999) and Weigmann et al. (2003). For each 2 h interval I will list the corresponding morphological stages[1] described in Campos-Ortega and Hartenstein (1997).

0–2 h: Fertilisation and the syncytium (morphological stages 1–4)

In the first 2 h after fertilisation, the nucleus of the *D. melanogaster* embryo goes through a series of 13 mitotic cell divisions. The time taken for each cell division increases with every cycle, with cycle 13 requiring approximately 21 min (Foe and Alberts, 1983). Only the nuclei are duplicated at each of these cycles, which are all contained inside a single shared cytoplasm. This body is called the

[1] In this work, I will use the term "morphological stage" to refer to a stage defined by the morphology of the embryo and the term "stage" and "developmental stage" to refer to a stage defined by the time since fertilisation.

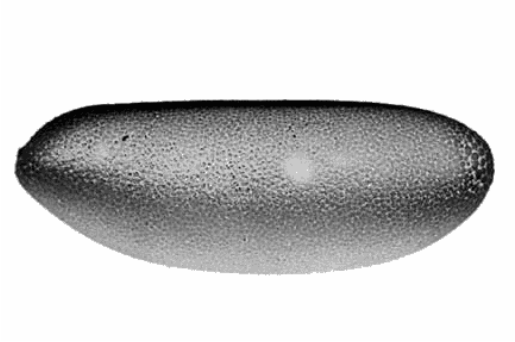


Figure 1.1.: *Drosophila melanogaster* embryo, lateral view, morphological stage 4. Colours inverted, microscopy by Dr. F. Rudolf Turner, Indiana University. Used with permission.

syncytium. After the fifth division, the nuclei move toward the periphery of the shared cytoplasm, where they form the syncytial blastoderm (Figure 1.1).

Patterning of the *Drosophila* embryo occurs within these first 3 h of development (St Johnston and Nüsslein-Volhard, 1992), and takes advantage of the fact that the syncytium allows for the free diffusion of molecules along the embryo. The shared cytoplasm allows maternal morphogens such as the proteins Bicoid and Nanos to form gradients, with Bicoid diffusing from the anterior (future location of head) and Nanos diffusing from the posterior (future location of tail). These gradients enable the establishment of the anteroposterior axis by differential regulation of transcription factors (see Section 1.3.1) such as Hunchback, Krüppel and Knirps (St Johnston and Nüsslein-Volhard, 1992). In addition, local activation of the transmembrane receptors Toll (Anderson et al., 1985) and Torso (Casanova and Struhl, 1989) help define the terminal areas at the anterior and posterior end as well as the dorsoventral axis, which spans from the back to the belly.

In this early phase of development, the vast majority of transcripts in the embryo come from the mother. Only around the 11^{th} cycle of cell division does widespread transcription of zygotic genes begin in what is called the maternal-to-zygotic transition (MZT) (Edgar and Schubiger, 1986). This activation of zygotic genes is linked to the rapid degradation of maternal RNA, involving both maternal RNA-binding proteins (Benoit et al., 2009) as well as zygotic non-coding RNA (see Section 1.3.3). The MZT is completed with the midblastula transition (MBT) after the 13^{th} cell cycle, when zygotic gene products are required for development to proceed.

2–4 h after fertilisation: Cellularisation and gastrulation (morphological stages 5–9)

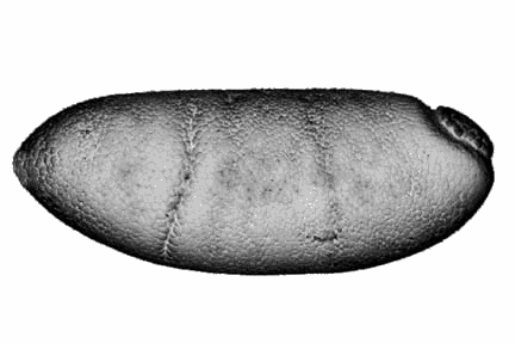


Figure 1.2.: *Drosophila melanogaster* embryo, lateral view, morphological stage 6. Colours inverted, microscopy by Dr. F. Rudolf Turner, Indiana University. Used with permission.

2–4 h after fertilisation is the first time point which I will be investigating in this study. It represents a very early stage in *D. melanogaster* embryo development, covering the transition from the syncytial blastoderm to the cellular blastoderm followed by gastrulation.

At 2 h 10 min after fertilisation and 13 cycles of mitosis, cellularisation of the embryo begins. The plasma membrane of the embryo grows inward, engulfing each individual syncytial nucleus to form a cellular blastoderm.

This process is then followed at 2 h 50 min by major cell shape changes and movements which mark the beginning of gastrulation (Figure 1.2) (Leptin, 1999). This process will separate the embryo into its three germ layers — the endoderm, mesoderm and ectoderm. The endoderm, starting from the terminal ends at the anterior and posterior, will give rise to the midgut. The mesoderm, which forms from ventral cells that invaginate inwards, will give rise to, among others, the muscles and the fat body. The ectoderm will give rise to the nervous system, epidermis, fore- and hindgut, the trachea and more.

Together, the ectoderm and mesoderm make up the germ band, with the ectoderm on the outside and the mesoderm on the inside. Starting from the 3 h 10 min mark, this germ band quickly elongates, folding back upon itself on the dorsal side.

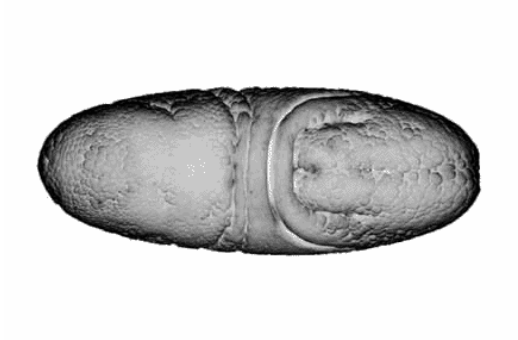


Figure 1.3.: *Drosophila melanogaster* embryo, dorsal view, morphological stage 9. Colours inverted, microscopy by Dr. F. Rudolf Turner, Indiana University. Used with permission.

4–6 h after fertilisation: Germ band elongation (morphological stages 9–11)

During this time point, the germ band elongates further (Figure 1.3), but more slowly. It reaches its maximum length at approximately 5 h after fertilisation, having folded back upon itself for about ¾ of the embryo. Neuroblasts, which form from the neurogenic region of the ectoderm, begin to divide, forming ganglion mother cells which will give rise to the central nervous system.

6–8 h after fertilisation: The extended germ band (morphological stages 11–12)

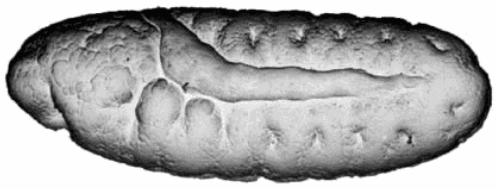


Figure 1.4.: *Drosophila melanogaster* embryo, lateral view, morphological stage 11. Colours inverted, microscopy by Dr. F. Rudolf Turner, Indiana University. Used with permission.

This is the second time point I am investigating in this study. During this middle stage of embryogenesis, the cells arrange to form more clearly visible segments

(Figure 1.4). These segments, delineated by parasegmental furrows, will give rise to different parts of the adult fly.

At 7 h 20 min after fertilisation, the germ band begins to retract. Around this time, cells in specific locations start to undergo programmed cell death. This phenomenon of deliberate, coordinated removal of cells will continue to occur in different parts of the embryo throughout its development (Abrams et al., 1993).

8–10 h after fertilisation: Germ band retraction (morphological stages 12–13)

Figure 1.5.: *Drosophila melanogaster* embryo, lateral view, morphological stage 12. Colours inverted, microscopy by Dr. F. Rudolf Turner, Indiana University. Used with permission.

The main event during this time point is the continued retraction of the germ band (Figure 1.5), which finishes at approximately 9 h 20 min after fertilisation. In the process of this shortening, the segments of the germ band also become more clearly visible. At the same time, the anterior and posterior midgut extend toward each other, until they meet in the middle of the embryo.

10–12 h after fertilisation: Tissue differentiation and dorsal closure (morphological stages 13–15)

10–12 h after fertilisation is the last time point that I am investigating in this study, covering the end of germ band retraction and the beginning of cell differentiation.

After retraction of the germ band, organ precursor cells (primordia) begin to express cell-type specific markers and differentiate. The segments of the germ band are now clearly separated into 12 parts, with segments T1–3 making up the future thorax and segments A1–9 forming the abdomen (Figure 1.6). This results in a mix of cells that are now specialising, with large cell-type specific differences.

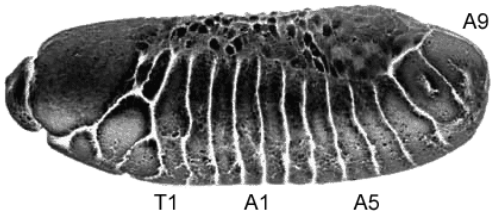


Figure 1.6.: *Drosophila melanogaster* embryo, lateral view, morphological stage 13. Colours inverted, microscopy by Dr. F. Rudolf Turner, Indiana University. Used with permission.

The type of body part that these segments develop into are controlled by the Hox (homeobox-containing) transcriptional regulators (see Section 1.3.1). These eight DNA-binding proteins are split up into the Bithorax complex, which controls the differences between abdominal and thoracic segments, and the Antennapedia complex, which controls the differences between thoracic and head segments (McGinnis et al., 1984). Mutations in these genes can have major effects on the adult body structure. For example, Ultrabithorax is responsible for regulating the differences between the T2 and T3 segments (Struhl, 1982; Weatherbee et al., 1998). A loss of this gene will result in the T3 segment developing like T2, producing a second pair of wings instead of halteres. On the other hand, a mutant with a gain of function of this gene in T2 will grow into a wingless fly with two pairs of halteres.

At 10 h 20 min after fertilisation, the head structures begin to move into the interior of the embryo in a process called head involution. Dorsal closure begins after approximately 11 h after fertilisation. During this process, the hole that has been left in the dorsal epithelium by the retraction of the germ band is closed by lateral epithelium, which is coming up from both sides of the embryo and merges at the dorsal midline. This is the last major morphogenetic movement of *Drosophila* embryogenesis.

12–22 h after fertilisation: End of embryogenesis (morphological stages 15–17)

Dorsal closure completes around 13 h after fertilisation (Figure 1.7), followed by the completion of head involution. At the same time, the outer layer of the larva,

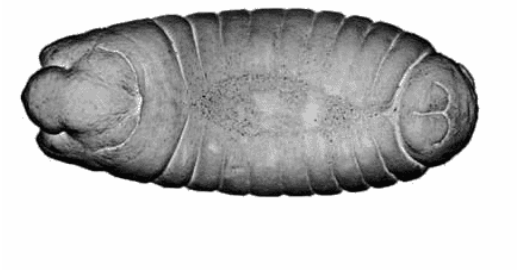


Figure 1.7.: *Drosophila melanogaster* embryo, dorsal view, morphological stage 15. Colours inverted, microscopy by Dr. F. Rudolf Turner, Indiana University. Used with permission.

the cuticle, begins to form. This layer protects the larva from the outside and is required for its structural integrity (Ostrowski et al., 2002). Approximately 21–22 h after fertilisation, the larva hatches.

1.3. Gene regulation

In order to provide proteins and other products required by an organism, genes are transcribed into complementary molecules of RNA, which can act as both carriers of information for further processing as well as functional molecules themselves. Transcription of a gene begins at the transcription start site (TSS) at its $5'$ end and then proceeds towards its $3'$ end. Many genes give rise to messenger RNA (mRNA), which will be translated by ribosomes to produce proteins. In addition, there are also genes that are transcribed into non-coding RNA (ncRNA), which may be further processed but not translated. This process of reading the information contained in a gene to synthesise a functional product is called gene expression. The amount of functional product that is being produced can differ from gene to gene, resulting in different levels of gene expression. The process that integrates information from genetic features and other signals to give rise to these differences is called gene regulation.

While all of the cells in an organism contain the same DNA (except in unusual circumstances) and thus the same genes, the expression level of a gene can also vary greatly between different cells or conditions. This is possible because gene regulation can be affected by a variety of different factors, including the presence of certain kinds of proteins, cell-cell interactions or environmental stimuli. Cru-

cially, regulatory changes can also be passed on to the mitotic offspring of cells, through processes such as the auto-regulation of transcription factors (Harding et al., 1989), the transmission of structural DNA features (Ringrose and Paro, 2004) or non-coding RNA (Pauli et al., 2011). Thus, a cell lineage can become committed to a certain developmental programme and pass this programme on to its offspring, even when the original stimulus is no longer present. During embryogenesis it is these processes that establish the different cell types in the organism.

The expression of a protein-coding gene can be regulated at each of the steps from the DNA to the fully functional product. In eukaryotes, the major levels of this process, shown in Figure 1.8, are: (1) transcription of DNA into precursor-mRNA (pre-mRNA); (2) processing of pre-mRNA into capped, spliced, polyadenylated mRNA and its export into the cytosol; (3) post-transcriptional degradation of mRNA; (4) translation of mRNA into protein; and (5) post-translational modification of proteins.

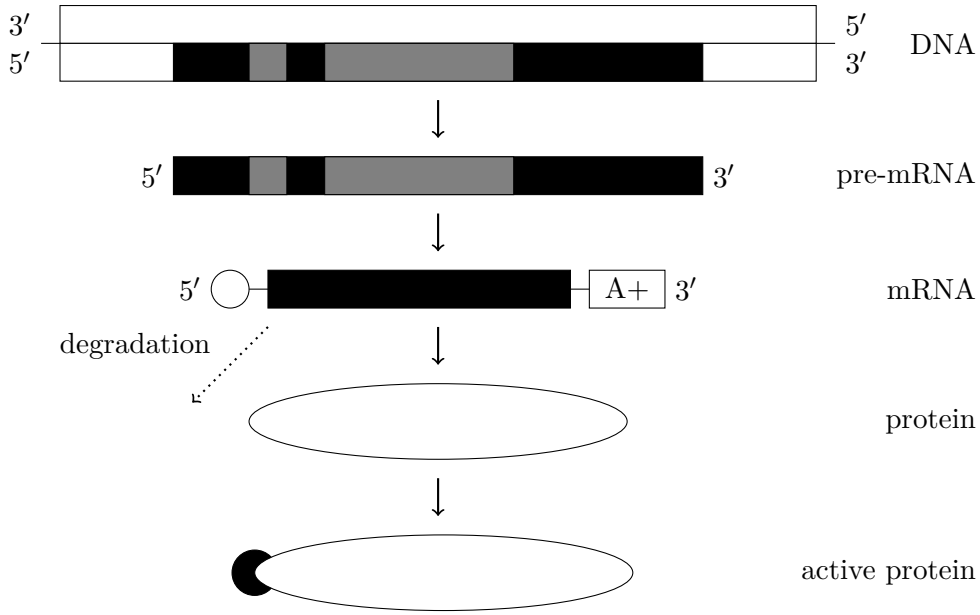


Figure 1.8.: Schematic of the path from DNA to an active protein. Region of DNA shown in white, exons of a gene shown in black, introns shown in grey, 5′ cap as a white circle. A+, poly(A) tail.

Each of these steps can be sped up, slowed down or completely inhibited by regulatory mechanisms. The first step, transcription, can be regulated by affecting the function of RNA polymerase, which transcribes DNA into RNA. An exam-

ple of this is the pair-rule gene *even-skipped* (*eve*), which is expressed in seven stripes along the anteroposterior axis (Nüsslein-Volhard and Wieschaus, 1980). The expression of each stripe is constrained to a small subset of cells in the embryo, based on the gradients of multiple transcriptional regulators such as Bicoid, Hunchback, Krüppel and Giant (Small et al., 1992).

At the level of processing, the protein-coding sequence of mRNA can be altered through, among others, the regulation of splicing or changes to the location of its 3′ end. The regulation of the gene *e(r)*, which is important in the female germ line of *Drosophila*, is an example of such regulation through RNA processing (Gawande et al., 2006). I will expand on this example in Section 1.3.2.

By regulating the translation of mRNA by ribosomes, the amount of protein that is produced from a molecule of mRNA can be modified. For example, translation of mRNA from the gene *oskar*, which is a maternal mRNA important for posterior body patterning in *Drosophila*, is repressed by the protein Bruno (*arrest*), ensuring that Oskar is only produced at the posterior pole of the embryo (Kim-Ha et al., 1995).

In addition, the sequence-specific binding of ncRNAs has also been shown to repress translation of mRNAs as well as promote their degradation. For example, the *bantam* gene in *Drosophila* codes for a ncRNA that can target a region in the 3′ untranslated region (3′ UTR) of the pro-apoptotic gene *hid*, downregulating its expression and thus preventing apoptosis (Brennecke et al., 2003).

Finally, proteins can be modified in various ways after they have been translated, which can affect or even be required for their function. A classic example of this from *Drosophila* is the phosphorylation of Period, which undergoes a daily cycle and is important for maintaining the circadian rhythm (Edery et al., 1994).

Although regulation of translation and post-translational modification are important processes, gene expression is often measured at the level of steady-state mRNA concentration, as the quantification of protein concentrations is significantly more complex (Vogel and Marcotte, 2012; Csárdi et al., 2015). In line with this, I will be using steady-state mRNA levels estimated using the 3′ Tag-Seq high-throughput sequencing protocol to determine gene expression levels for this project. In the following section, I will describe in more detail the major ways in which gene regulation can affect these steady-state levels of mRNA.

1.3.1. Transcriptional regulation

In eukaryotes, all pre-mRNAs as well as many ncRNAs are transcribed by the second RNA polymerase, RNA Pol II (reviewed in Fuda et al., 2009). With the aid of other proteins, called general transcription factors, RNA Pol II binds to the core promoter of a gene to begin the process of transcription (Smale and Kadonaga, 2003). Once it has bound to the promoter, RNA Pol II must undergo a conformational change which releases it from the general transcription factors and allows it to start moving along the DNA, producing a complementary molecule of RNA (Phatnani and Greenleaf, 2006). This elongation phase of transcription can be further regulated through processes such as RNA Pol II pausing (reviewed in Zhou et al., 2012).

Many parts of this process can be either supported or hindered by different factors, which are called transcriptional regulators or transcription factors. These regulators are mainly DNA-binding proteins which recognise specific motifs in regulatory regions, usually approximately 6–12 bp in size (Spitz and Furlong, 2012). The regions bound by transcriptional regulators are called *cis* regulatory modules (CRMs). The term *cis* indicates that the regulator binds the same molecule of DNA from which the gene will be transcribed. A regulator that can affect the expression of genes on different molecules of DNA is said to be acting in *trans*.

Unlike in prokaryotes, where CRMs are usually located close to the promoter region, eukaryotic CRMs can be tens and even hundreds of kilobases away from the promoter, even inside the gene itself (Bulger and Groudine, 2011). This is made possible by the formation of DNA loops, which can bring the CRMs physically close to the promoter region in 3D space, even when they are located far away on the linear chromosome (Ghavi-Helm et al., 2014). Genes can have many different CRMs, which will all act in concert to determine the level of transcription, integrating the signals from multiple individual transcriptional regulators (Spitz and Furlong, 2012). At the same time, a single CRM can also affect the expression of multiple genes (Link et al., 2013).

There are two main kinds of transcriptional regulators — activators and repressors. Some of them act directly on RNA Pol II or its transcription factors, but most of them recruit secondary proteins called co-activators and co-repressors (such as the Mediator complex, see Conaway and Conaway, 2011), which then either activate or repress transcription through further interactions.

Activators increase the level of transcription, often by recruiting Pol II to the promoter, or by releasing a Pol II that has paused at the promoter or further along

the gene body. Repressors decrease the level of transcription, by either hindering the action of activators, or interacting with the general transcription factors.

Many regulatory sequences have been identified through reporter assays, in which a putative regulatory region is cloned in front of a reporter gene (Bulger and Groudine, 2011). If the levels of the reporter gene are higher when the sequence is present than in a control, this sequence is called an enhancer. If they are lower, the sequence is called a silencer. Enhancer and silencers are likely to contain CRMs bound by an activator or repressor respectively, but as they are experimentally defined, the exact mechanism and location are not necessarily known. The first enhancer was described in 1981, a 72 bp repeat sequence from simian virus 40 (SV40) that was found to increase the expression level of the *β-globin* gene in *cis* (Banerji et al., 1981).

The binding affinity of RNA Pol II, as well as transcriptional regulators, is also associated with the accessibility of the DNA (Knezetic and Luse, 1986). Eukaryotic DNA is usually tightly packed in a complex called chromatin, with DNA wrapping around histones to form structures called nucleosomes (Felsenfeld and Groudine, 2003). These nucleosomes allow for an efficient packing of the DNA, but also make the DNA less accessible to RNA Pol II and other DNA-binding proteins, influencing the rate of transcription. Chromatin structure can be changed through a variety of ways, including by transcriptional regulators called pioneer factors, which can alter the accessibility of chromatin and recruit downstream regulators (Magnani et al., 2011). In addition, epigenetic modifications such as the acetylation and methylation of histones (Bannister and Kouzarides, 2011) are commonly observed in regions associated with expressed genes, such as active promoters (Barski et al., 2007). However, whether these epigenetic modifications actually affect transcription or are merely a symptom of it remains controversial (Ptashne, 2013).

1.3.2. Regulation of RNA processing

After transcription, the pre-mRNA of eukaryotic protein-coding genes still needs to be processed into mature mRNA and exported to the cytoplasm before it can be translated into protein. These processing steps often happen cotranscriptionally, while the RNA is still being transcribed by *RNA Pol II* (reviewed in Proudfoot et al., 2002; Moore and Proudfoot, 2009).

An important mRNA processing step in eukaryotes is splicing, during which non-coding regions of the pre-mRNA (introns) are excised, leaving only the cod-

ing regions (exons) (Padgett et al., 1986). Introns almost always contain the dinucleotides GU and AG at their 5′ and 3′ splice sites, respectively. The process of splicing starts with the formation of a 5′ to 2′ bond between a specific A nucleotide near the 3′ end of the intron and the 5′ splice site. This cuts the bond between the 5′ exon and the 5′ end of the intron, and joins the intron to itself forming a lariat (loop). The released 3′ end of this exon then forms a new bond with the exon at the 3′ splice site, joining the two exons and releasing the intron lariat. The inclusion or exclusion of individual exons during splicing, a process called alternative splicing, is a major gene regulatory mechanism in eukaryotes, and allows a single gene to give rise to many different transcript isoforms (Shin and Manley, 2004; Kornblihtt et al., 2013).

In addition, each eukaryotic mRNA has a cap added to its 5′ end and a tail of multiple A nucleotides (poly(A) tail) added to its 3′ end, which mark the mRNA as a functional transcript and prevent its degradation (see Section 1.3.3). Only complete, successfully spliced mRNAs are allowed to leave the nucleus and move to the cytosol, where they will be translated into proteins (Stutz and Izaurralde, 2003).

Particularly relevant to the experimental design of my study is the addition of the poly(A) tail (reviewed in Proudfoot, 2011). The two main proteins involved in this process are called CstF (cleavage stimulation factor) and CPSF (cleavage and polyadenylation specificity factor). These two proteins are carried on the tail of RNA Pol II, allowing them to read the RNA nucleotides as they are being transcribed from the DNA. CPSF recognises the canonical poly(A) motif AAUAAA[2], which is located approximately 10–30 bp upstream of the cleavage site at the 3′ end of the transcript. At the same time, CstF binds to a GU- or U-rich region located up to 30 bp downstream of the 3′ end, which is called the downstream sequence element (DSE). Once these two proteins are bound, they recruit additional cleavage factors, which cleave the transcript at the 3′ end.

After cleavage, PAP (poly-A polymerase) is recruited to add A nucleotides to the 3′ end of the transcript, forming the poly(A) tail. Crucially, the poly(A) tail is not encoded on the DNA but added by PAP without a template. The length of this poly(A) tail can vary between genes and conditions, but is usually around 200–300 bp long (Colgan and Manley, 1997).

Differences in the composition of the DSE or a region upstream of the canonical motif called the upstream sequence element (USE) have been shown to modulate

[2]In this dissertation I also refer to this motif by its DNA equivalent, AATAAA

the affinity of the poly(A) machinery to the poly(A) site (see for example Gil and Proudfoot, 1987; Carswell and Alwine, 1989). Consequently, a single gene can have multiple poly(A) sites of varying strength, each of which has a certain chance of terminating transcription every time a new molecule of RNA is transcribed. If a poly(A) site is weak, CPSF and CstF will often fail to bind to it, resulting in Pol II "reading through" to the next poly(A) site. This will result in the transcription of two different 3′ isoforms of the mRNA at different levels, one shorter and one longer. Such alternate use of poly(A) sites is called alternative polyadenylation (APA, Di Giammartino et al., 2011).

Usually, APA does not affect coding regions of genes and the different transcript isoforms will only differ in the length of their 3′ UTR. This can be used to regulate the steady-state levels of RNA, as longer 3′ UTRs may contain more miRNA binding sites, which can have a large effect on mRNA stability (see Section 1.3.3). However, there are also cases where APA is associated with alternative splicing and results in the inclusion or omission of part of the protein structure of a gene. An example of this is the *IgM* gene in humans, where the use of alternate poly(A) sites results in either a membrane-bound or a secreted protein isoform (Takagaki et al., 1996).

In *Drosophila*, an example of APA is the expression of a sex-specific isoform of the *e(r)* gene in the female germ line (Gawande et al., 2006). A schematic of this gene and its two primary poly(A) sites is shown in Figure 1.9. The promoter-proximal poly(A) site uses a weak version of the canonical poly(A) signal, in which the first A nucleotide is replaced by a T. The second poly(A) site, located 221 bp downstream, uses the exact canonical signal, and thus has a higher affinity for the poly(A) machinery. The proximal poly(A) site is followed by a GU-rich DSE region, which allows CstF-64 (a component of the CstF complex) to bind. In males, the proximal poly(A) site is used exclusively, suggesting that the weak poly(A) site together with the binding of CstF-64 enables strong polyadenylation.

This poly(A) site usage can be switched by the product of the *Sex-lethal* (*Sxl*) gene, which is crucial for sex determination (Samuels et al., 1991). The functional isoform of this splicing regulator, which can bind to U-rich sequences in RNA, is only expressed in female individuals. In the female germ line, this regulator binds to the GU-rich element downstream of the proximal poly(A) site of *e(r)*. There it competes for binding with CstF-64, decreasing the affinity of the poly(A) machinery to this proximal poly(A) site. As polyadenylation is thus prevented from occurring at the proximal poly(A) site, this interaction results in an increase

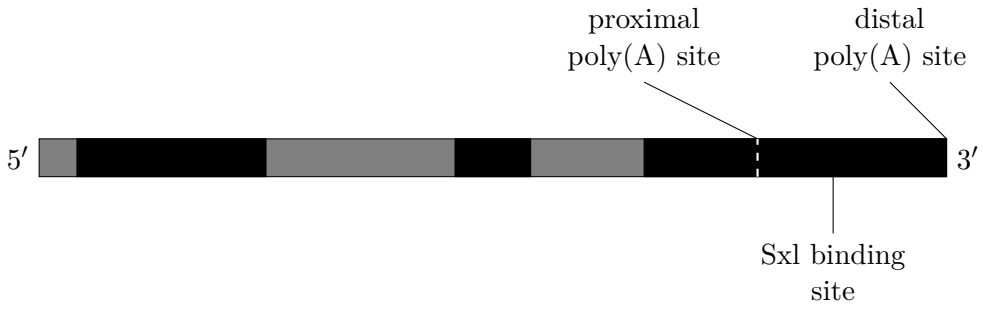


Figure 1.9.: Schematic of the gene structure of *e(r)*, not drawn to scale. Exons shown in black, introns shown in grey.

in the production of transcripts ending at the distal poly(A) site.

APA has also been shown to occur transcriptome-wide, with 3′ UTRs globally increasing in length during mouse development. It has been suggested that this is a deliberate process, allowing for increasingly fine-grained post-transcriptional control through regulators such as miRNA as development progresses (Ji et al., 2009).

1.3.3. Post-transcriptional regulation

mRNAs lacking the 5′ cap or the 3′ poly(A) tail are rapidly degraded by the cell (reviewed in Parker and Song, 2004). This process serves a variety of purposes, including quality control of mRNA (Maquat and Carmichael, 2001), removal of side-products of transcription such as debranched spliced introns, and removal of foreign RNA (Anderson and Parker, 1998). As the poly(A) tail of mRNA is gradually shortening during its lifetime, the degradation machinery also ensures the eventual degradation of all mRNAs, with the speed of degradation dependent on the length of their poly(A) tail (Decker and Parker, 1993).

In addition, over the last few years, more and more non-coding RNAs (ncRNAs) have been discovered, which can repress the translation and initiate degradation of mRNA in a sequence-specific manner in a process called RNA interference (RNAi). The types of ncRNAs that are involved in this process include microRNAs (miRNAs), small interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs) (Valencia-Sanchez et al., 2006; Ghildiyal and Zamore, 2009; Jonas and Izaurralde, 2015). Some of these ncRNAs have been shown to be key regulators during animal development (Stefani and Slack, 2008).

When miRNAs and other ncRNAs associate with a member of the Argonaute

protein family, they form a RNA-induced silencing complex (RISC). The RISC is guided to the target mRNA through complementary base-pairing of the ncRNA sequence with the mRNA sequence. The binding of RISC to the target mRNA then leads to downregulation of the gene, through repression of translation and/or cleavage of the mRNA, which is then degraded through the normal mRNA degradation pathways (Djuranovic et al., 2012).

The regions bound by miRNAs are often located in the 3′ UTR of mRNA (Bartel, 2009). For example, the degradation of maternal mRNAs in the MZT has been associated with zygotic miRNAs binding to the 3′ UTR of maternal mRNAs (Bushati et al., 2008). Thus, a change in the length of the 3′ UTR region, caused by APA, can result in an increase or decrease in regulation by miRNAs.

1.4. Estimation of gene expression levels with RNA sequencing

Early methods to estimate the RNA levels of genes were the Northern blot (Alwine et al., 1977) and quantitative reverse transcription polymerase chain reaction, qRT-PCR (Gibson et al., 1996). In a Northern blot, RNA is separated by gel electrophoresis and then visualised by hybridisation with labelled probes. In qRT-PCR, RNA is reverse transcribed into complementary DNA (cDNA) using a reverse transcriptase and then amplified using PCR, after each cycle of which the concentration of DNA is measured using a fluorescent dye. However, both of these methods are very low-throughput, not very accurate and would require large amounts of starting material to estimate the expression level of all the genes expressed in a higher organism.

In 1995, a method for estimating the expression levels of many genes simultaneously using DNA microarrays was introduced (Schena et al., 1995). Like qRT-PCR this method relies on the reverse transcription of RNA into cDNA. This cDNA is then labelled with a fluorescent dye and hybridised to a DNA microarray containing complementary DNA for thousands of known transcripts at known locations. The RNA levels can then be estimated by measuring the intensity of fluorescence at each location and either normalising it using spike-ins of known concentration or directly comparing it to a second sample on the same microarray using two different fluorescent dyes. Towards the end of the 20$^{\text{th}}$ century, microarrays became the most commonly used method of measuring gene

expression levels. However, since microarrays only allow the measurement of RNA for which the sequence is known, they are not suitable for the detection of novel transcripts or novel splice isoforms. They are also often unable to measure the expression of transcripts with low abundance due to background noise (Gautier et al., 2004) and are not necessarily suitable for studying the absolute expression of genes in a single sample (Allison et al., 2006).

Recent improvements to high-throughput sequencing now allow for the direct sequencing and quantification of cDNA libraries (Mortazavi et al., 2008). This method, called RNA-seq, has since been shown to be superior to microarrays in almost all regards except cost (Marioni et al., 2008). In addition to the mere quantification of known transcripts, RNA-seq also enables the discovery of new transcript isoforms or entirely unknown genes.

1.4.1. Standard poly(A)$^+$ RNA-seq

In standard poly(A)$^+$ RNA-seq (Mortazavi et al., 2008), polyadenylated RNA is captured using an oligo-dT primer that will bind to the complementary poly(A) tail. This selection for polyadenylated RNA is performed to increase the fraction of mRNA (which are usually the molecules of interest) in the overall sample. Without this step, any signal generated by the mRNAs in the sample would be overshadowed by the large amounts of ribosomal RNA (rRNA) present in every cell, which accounts for most of the cell's total RNA.

Once captured, the polyadenylated RNA is fragmented into parts of approximately 200–300 bp in size. These fragments are then reverse transcribed into cDNA and sequenced using high-throughput short-read sequencing. Before sequencing, the cDNA is usually amplified using PCR to allow quantification even from small starting quantities of RNA. A diagram of RNA-seq is shown in Figure 1.10.

The data generated by RNA-seq consists of short (often approximately 100 bp) reads giving the sequence of either one end (single-end) or both ends (paired-end) of the RNA fragments. Depending on the exact protocol used, reads will either always be generated from the same or opposite strand as the RNA (stranded) or from a random strand (unstranded). When mapped back to the reference genome of the organism, these reads will cover all expressed exons of transcripts, since fragmentation occurs at random. Reads that span a splice junction between two exons will contain a gap when mapped to the genome, which makes it possible to identify introns. In addition, alternative splicing can be observed by comparing

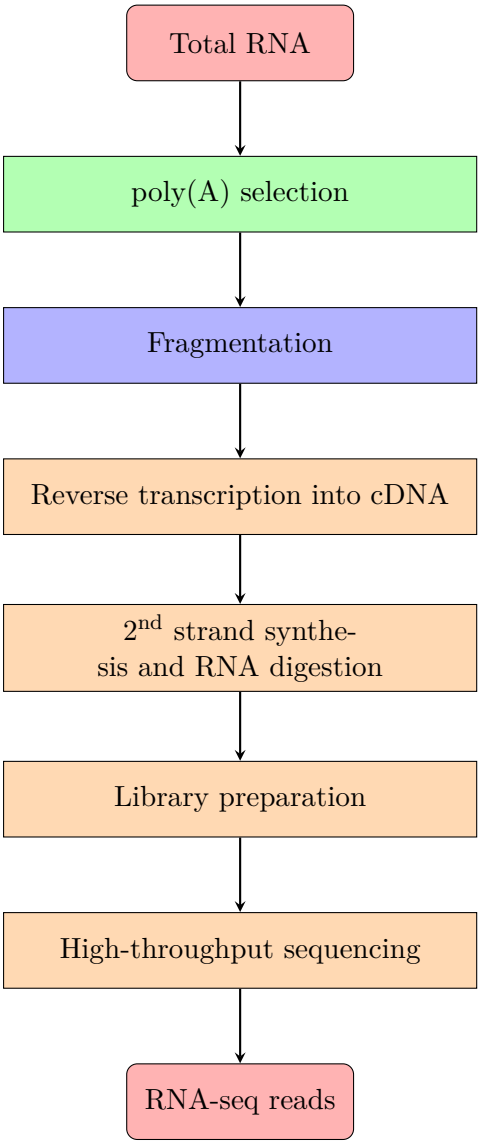


Figure 1.10.: The poly(A)$^+$ RNA-seq protocol.

the number of reads mapping to different exons of the gene.

As the number of reads per annotated transcript will be approximately proportional to the number of RNA fragments present in the sample, the count of reads mapped to a transcript can be used to estimate its expression level. For comparisons between genes, this count needs to be normalised by the length of the transcript to account for the number of fragments generated from a single molecule of RNA. In addition, the count is usually normalised by the overall number of sequencing reads mapped to the genome, to allow for comparisons

between samples with different numbers of sequenced reads (sequencing depth). The expression level of a transcript, after accounting for these factors, is usually expressed in fragments per kilobase of transcript per million reads mapped (FPKM) or transcripts per million (TPM).

1.4.2. 3′ Tag-Seq

In this study, I use a different but largely similar protocol to standard RNA-seq, called 3′ Tag-Seq (Yoon and Brem, 2010; Wilkening et al., 2013). Instead of sequencing whole transcripts as in RNA-seq, in 3′ Tag-Seq only fragments ending with the 3′ poly(A) tail (the 3′ tags) are sequenced. This is achieved by first fragmenting the RNA and then performing reverse transcription using an anchored oligo-dT primer, which will produce cDNA only for fragments that end in poly(A). Thus, only the 3′ ends of polyadenylated RNAs will be available for sequencing and each molecule of RNA will only yield a single fragment. A diagram of 3′ Tag-Seq is shown in Figure 1.11.

In the variant of 3′ Tag-Seq that I will be using for my project, each of these cDNA fragments is sequenced using stranded, single-end reads starting from the 5′ end of the fragment, towards the 3′ end. When mapped to the genome, these reads will not cover the entire transcript, but will instead be concentrated at the 3′ end of transcripts, forming a peak shape. As each sequencing read can be assumed to represent one molecule of RNA, the number of reads mapped to this peak region (or the height of the peak) can be used to estimate the expression level of the transcript, without normalising for transcript length. However, as in standard RNA-seq, normalisation by the total number of mapped reads is still required to account for different sequencing depths between samples.

In addition to being suitable for the determination of overall gene expression levels, 3′ Tag-Seq also provides an additional level of detail over standard RNA-seq, as different 3′ poly(A) sites can be identified as separate peaks of 3′ Tag-Seq reads. This is not always possible with RNA-seq, where signal coming from a short transcript isoform is difficult to distinguish from signal coming from a longer one. This feature of 3′ Tag-Seq and its variants (e.g. 3P-seq, see Jan et al., 2011) has been used to perform genome-wide screens of polyadenylation sites in organisms such as zebrafish (Ulitsky et al., 2012), yeast (Wilkening et al., 2013) and *Drosophila* (Smibert et al., 2012). On the other hand, as reads are obtained only from the 3′ ends of transcripts, 3′ Tag-Seq cannot provide any information about alternative splicing of internal exons.

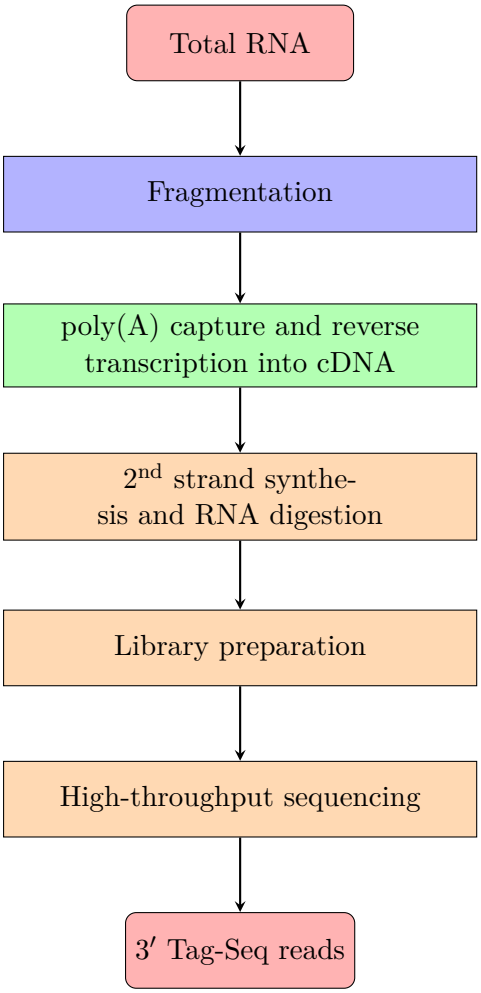


Figure 1.11.: The 3′ Tag-Seq protocol.

An important source of artefacts in 3′ Tag-Seq is the potential for the oligo-dT primer to capture fragments of RNA that contain a poly(A) sequence, but are not actually real poly(A) tails (Nam et al., 2002). If not filtered out, the reads generated by these segments can form a shape just like a real 3′ end peak, resulting in the identification of false positive 3′ ends. In Chapter 2, I will describe how I accounted for this problem.

1.5. Genomic variation

When Darwin proposed his theory of evolution, it was unclear which processes gave rise to variation in populations. Today we know of course that the main

driver of biological variation is mutation of DNA. Mutations can arise spontaneously, but can also be induced through environmental factors, such as exposure to radiation (Muller, 1927). A mutation that affects a cell of the germ line will be passed on to the offspring, resulting in inheritance of variation.

Locations in the genome that differ between individuals in a population are called polymorphisms or variants and the different stretches of DNA associated with them are called alleles. Nearly all animals are diploid, which means that they carry two sets of homologous chromosomes. The collection of homologous alleles that an individual carries comprises its genotype and is usually written as a string of two letters. If both copies of the allele in an individual are the same, the individual is said to be homozygous for the polymorphism, otherwise it is said to be heterozygous. For example, an individual with genotype AA is homozygous for (carries two copies of) allele A, while an individual with genotype Aa is heterozygous for alleles A and a. The allele that is less common in the population is called the minor allele, and its frequency is the minor allele frequency (MAF).

The most common type of polymorphism is the single-nucleotide polymorphism (SNP), in which a single base pair of DNA differs between individuals in a population (The 1000 Genomes Project Consortium, 2012). Despite such a small change, a difference of a single nucleotide can still have a variety of effects. For example, polymorphisms in sequences recognised by DNA- or RNA-binding proteins can cause a decrease or increase in the binding proteins' affinity, which can have various effects on the regulation of gene expression (see Section 1.3). In addition, changes to the coding region of genes can alter their amino acid composition, by changing the identity of a single amino acid or by causing translation of the transcript to terminate early or extend beyond the normal $3'$ end.

Most SNPs are biallelic, with only two possible alleles (and three possible genotypes) found in the population (The International HapMap Consortium, 2005). This is because two separate mutation events would have had to occur at the exact same location for more than two alleles to arise. While such polyallelic SNPs are known to exist, they are often more likely to be the result of genotyping errors rather than real biological variation (MacArthur et al., 2012).

In addition to changes of single nucleotides, stretches of DNA can also be inserted into or deleted from the genome (The 1000 Genomes Project Consortium, 2012). By convention, if these variants are 2–200 bp long they are called insertions/deletions (indels). Like SNPs, indels can have various effects on the

organism. For example, they can result in the deletion or introduction of an entire regulatory region, which can have a major effect on the regulatory landscape (Mullaney et al., 2010). While indels are usually not long enough to completely delete a gene from the genome, they can essentially disable it by removing its core promoter element or parts of the coding sequence (MacArthur et al., 2012). In particular, an insertion or deletion in the coding region of a gene with a length not divisible by three will change the interpretation of all following DNA triplets (codons) during translation, as it causes a shift in the reading frame. Such frameshift mutations almost always result in a non-functional protein with an altered amino acid structure and an abnormal length.

Longer insertions or deletions, as well as other structural rearrangements such as duplications, are called structural variation (SVs, Feuk et al., 2006). A type of SV that has played a major role in the study of *Drosophila* is the inversion, which can arise when a chromosome breaks and subsequently reassembles, with some of the fragments having become inverted. Inversions can be visualised on the oversized polytene chromosomes of salivary glands in *Drosophila*, allowing for their direct observation under the microscope (Painter, 1933). Since inverted regions of a chromosome are no longer homologous to their uninverted counterparts, recombination between the two alleles will be suppressed (Sturtevant, 1921). Because of this, inversions have played an important role in *Drosophila* mutagenesis screens, as chromosomes carrying inversions can be used as balancer chromosomes to maintain a heterozygous stock carrying a recessive mutation (Muller, 1927; Hentges and Justice, 2004).

1.6. Linkage and genetic association studies

A common task in genetics is to identify genetic differences between individuals that are associated with differences in their phenotypes. Historically, such questions have been answered by linkage mapping, as introduced by Sturtevant in 1913 (see Section 1.1.2). Briefly, this involves crossing lines carrying genetic markers with those carrying a phenotype of interest and then studying the frequency of cosegregation between each marker and the phenotype in the offspring. If the traits are on different chromosomes, this frequency is expected to follow Mendel's Law of Independent Assortment. However, if the genes underlying the traits are located on the same chromosome, their cosegregation rate will decrease with decreasing distance between them. Thus, by studying the cosegregation rate with

many different markers, a gene can be pinpointed to a small region of the genome. This concept was first applied to dichotomous traits but was soon also extended to quantitative traits. For example, in 1923, Karl Sax published a study of the linkage between seed colour, colour pattern and seed weight in genetic crosses of *Phaseolus vulgaris* (Sax, 1923).

Linkage mapping is often conducted in inbred strains of laboratory organisms. These inbred strains are usually generated through at least 20 generations of brother-sister mating, which results in a group of individuals that are genetically almost identical (isogenic) and near-homozygous at all loci in the genome. By further inbreeding, these strains can be maintained for many generations, providing a theoretically infinite stock of genetically identical individuals. This allows specific crosses to be set up that can be used to study individual genes in a targeted way.

In humans, such inbreeding and designed crosses are of course not possible. However, similar studies can be conducted by tracing the transmission of phenotypes as well as genetic markers through a family tree. Using naturally occurring DNA variation as genetic markers (Botstein et al., 1980), such pedigree studies facilitated the discovery of genetic factors underlying Mendelian diseases such as Huntington's disease (Gusella et al., 1983).

1.6.1. Genetic association studies

An alternative approach for studying common diseases is the genetic association study, which compares the frequency of genetic variants between individuals with and without a certain trait in a population (reviewed in Altshuler et al., 2008).

An important assumption behind association studies is that, given a freely mating population, the genotypes at the vast majority of variants will be randomly distributed with respect to the phenotype, even if they are located on the same chromosome as the causal variant. Only the genotypes of causal variants and variants very close to it are correlated with the phenotype. This is due to linkage disequilibrium (LD), the non-random distribution of genotypes of nearby variants between individuals due to a combination of the local recombination rate and the population history of these variants. When genotyping costs were high, a high degree of LD in a population was thus advantageous for an association study, as it decreased the number of variants that had to be genotyped. However, this came at the cost of being unable to fine map associations to single variants, as the strength of association between a phenotype and the causal variant can be

indistinguishable from the strength of association with other variants in LD.

There were some early successes using genetic association studies, including the identification of the ApoE-ϵ4 allele as a risk factor for Alzheimer's disease (Corder et al., 1993). However, early genetic association studies were limited to few variants in small candidate gene regions, due to the challenges involved in the identification and genotyping of individual variants. Identifying an appropriate candidate region was difficult, and such studies were also susceptible to false positives caused by population structure. Only when techniques such as DNA microarrays (Wang et al., 1998) enabled high-throughput genotyping of many variants at the same time, did it become possible to solve these limitations by extending genetic association studies to the whole genome.

Shortly after, efforts such as the International HapMap project (The International HapMap Consortium, 2005) showed that the genotypes of many variants in the human population are indeed strongly correlated with other nearby variants, confirming earlier assumptions about LD. Together, these variants form so-called haplotype blocks inside which recombination appears to have been limited. By selecting a representative variant from each haplotype block, a reduced set of variants that "tag" most of the diversity in a population can be defined. Thus, the number of variants that need to be genotyped and tested to survey the whole genome is greatly reduced. However, as described above, this high degree of LD also means that fine mapping of the exact causal locus is not always possible.

These developments started the age of genome-wide association studies (GWAS), in which thousands of variants across the entire genome were genotyped and tested for association in a population. The first GWAS in humans, on Age-Related Macular Degeneration (AMD), was published in 2005 (Klein et al., 2005). Consistent with previous studies, this GWAS identified a significant association between a mutation in the coding region of the *CFH* gene and AMD. Two years later the Wellcome Trust Case Control Consortium (WTCCC) published their landmark paper reporting results of GWAS on seven different diseases (The Wellcome Trust Case Control Consortium, 2007). Since then, many more GWAS have been conducted in humans, numbering more than 1,500 studies to date (Welter et al., 2014). In addition to dichotomous traits, GWAS have also been applied to find loci associated with quantitative traits (QTLs), for example height (Gudbjartsson et al., 2008) or blood pressure (Newton-Cheh et al., 2009).

While most large-scale GWAS to date have been conducted in humans, the same principles can also be applied to other organisms. Thus, over the last few

years, GWAS have also been performed in a large variety of other species, such as *Arabidopsis thaliana* (Atwell et al., 2010), rice (Huang et al., 2011) and mouse (Valdar et al., 2006). The populations used for such studies are often closed outbred populations, which are bred for maximum heterozygosity and genetic variability (Chia et al., 2005). However, despite major efforts to generate genetically diverse individuals, the degree of LD in closed, outbred populations can still be large and their variability may not reflect the true genetic diversity present in a wild population. Thus, it is challenging to fine map traits to causal loci and draw conclusions about the effects a polymorphism would have in the wild based on such a GWAS.

1.7. Statistics for association studies of quantitative traits

Both linkage mapping and association studies rely heavily on statistics to ascertain the significance of observations.

Ideally, the underlying model used in such a study should be able to capture all possible associations between all genotypes at a given variant and the phenotype. However, two common simplifications are often applied in large-scale trait association studies (reviewed in Bush and Moore, 2012, see for example Gudbjartsson et al., 2008, Newton-Cheh et al., 2009). First, only biallelic variants are tested, which means that only two different alleles and three different genotypes need to be considered. Secondly, variants are only tested for additive effects, without explicitly considering dominant or recessive effects. This means that each additional copy of the minor allele is considered to have the same effect, with the difference between the homozygous major and the heterozygous genotype being half the difference between the homozygous major and the homozygous minor genotype. In a QTL study this makes it possible to calculate, for a given variant, the minor allele count (the number of copies of the minor allele) in each individual i and then test for association between this value $x_i \in \langle 0, 1, 2 \rangle$ and a quantitative trait y_i. While this is only accurate for true additive effects, additive models have been shown to still be a good choice to detect dominant effects, although they are less powered to detect weak recessive effects (Lettre et al., 2007).

In a quantitative trait association study, one could test for such an association by calculating the Pearson correlation coefficient r between x_i and the phenotype y_i of each individual i and then testing the alternative hypothesis that $r \neq 0$ against the null hypothesis that $r = 0$. A low probability of observing the given

or a more extreme effect under the null hypothesis (a small p-value) would indicate that the count of the variant's minor allele was associated with the phenotype and the variant might thus be a QTL.

A more robust approach is to use the Spearman's rank correlation coefficient ρ instead (see for example Montgomery et al., 2010), which measures the correlation between the ranks of the data instead of the raw values. This makes it less sensitive to outliers and non-normality of the data than Pearson's correlation coefficient. Again, the alternative hypothesis that $\rho \neq 0$ can be tested against the null hypothesis that $\rho = 0$ to identify significant associations.

However, in practice there are usually several known and hidden confounders in any genetic association study, which need to be considered in the association test to prevent loss of power and false positives introduced by experimental noise. Thus, a more desirable option is to model the phenotype as a sum of a genetic and additional confounding effects using a linear mixed model (Yu et al., 2006; Zhang et al., 2010), for example:

$$\mathbf{y} \sim \boldsymbol{\mu} + \beta\mathbf{x} + \mathbf{c}_1 + \mathbf{c}_2 + \ldots + \boldsymbol{\psi} \tag{1.1}$$

Using this approach, one can model a vector of observed phenotypes $\mathbf{y}$ as the sum of the intercept vector $\boldsymbol{\mu}$, a genetic effect of the minor allele count $\mathbf{x}$ with effect size β, terms accounting for additional covariates $\mathbf{c}_i$ and the noise $\boldsymbol{\psi}$. The strength of association in such a model can be estimated by the p-value from a likelihood ratio test comparing a model with $\beta \neq 0$ against a model with $\beta = 0$.

Mixed models make it possible to model the genetic effect as a fixed effect (i.e. the effect of a factor which we have measured in each individual), while the covariates can be modelled as random effects. In contrast to fixed effects, random effects allow us to model the effects of factors which we could not measure in our samples. Instead, we account for them in the model by assuming that they come from a certain random distribution with a certain covariance structure, which we can typically estimate from the data. In this way, covariates such as the effect of population structure on gene expression can be integrated directly as a term into the model.

More recently, this univariate approach has also been extended to multivariate models, with which multiple, potentially correlated, phenotypes can be tested for association at the same time (Korte et al., 2012; Zhou and Stephens, 2014). This not only allows for the study of several related phenotypes, but also makes it

possible to test for association of a variant to a single phenotype under multiple conditions. Thus, one can distinguish variants that have an effect in all conditions from those that have an effect in only some conditions in a statistically rigorous manner. The LIMIX software package, which I will be using for this study, provides an efficient way of specifying and testing such multivariate linear mixed models (Lippert et al., 2014).

1.7.1. Multiple testing

A major consideration in GWAS is the number of false positive associations introduced by multiple testing. In a single association test, values of $p < 0.05$ or $p < 0.01$ are often considered significant associations, corresponding to a 5 % or 1 % probability of observing the given or a more extreme effect under the null hypothesis (of there being no association). However, if we apply this test to a million independent variants in a genome, we can expect to find $0.01 \cdot 10^6 = 10000$ false associations at a 1 % significance threshold, even if there is no true association.

There are several methods to account for this multiple testing problem. The most conservative is to apply a Bonferroni correction by simply dividing the significance threshold by the number of tests performed (Dunn, 1959). In our example, this would correspond to requiring $p < 0.01 \div 10^6 = 10^{-8}$. This method controls the familywise error rate (FWER), which is the probability of making one or more false discoveries. However, such a stringent threshold results in low statistical power, which is defined as the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true ($1 - \beta$, where β is the false negative or type II error rate). This means that a high number of samples is required to detect effects, especially when these effects are small.

Less conservative approaches aim to control the false discovery rate (FDR) instead of the FWER. The FDR represents the fraction of discoveries that are expected to be false positives. To illustrate, if we discover 500 significant associations at an FDR of 10 %, we can expect that 450 of these associations are true, while 50 are false positives. Thus, we can increase our statistical power to detect associations by controlling the FDR, but at the cost of a larger number of false positives. A commonly used method for controlling the FDR is the procedure of Benjamini & Hochberg (BH, Benjamini and Hochberg, 1995). This method can be used to calculate a BH-adjusted p-value, which represents the FDR at which this p-value would be rejected.

However, controlling the FDR with BH's method still limits the power to detect

associations and does not take into account possible correlations between tests (for example due to LD). Thus, the current best-practice method to adjust for multiple testing is to calculate empirical p-values through permutation experiments (Churchill and Doerge, 1994). In this method, the genome-wide search is not only performed on the real phenotype data, but also on n random permutations of the phenotype data, where n is usually in the order of thousands. For each random permutation, the best observed p-value (for what is almost certainly a false positive association) is recorded. The p-value obtained from the normal association test can then be ranked in this empirical null distribution to obtain an empirical p-value, which is corrected for multiple testing. Since this approach determines the null distribution based on the actual phenotypes and genotypes, it automatically accounts for the distribution of phenotypes (such as the possible presence of outliers) as well as genotypes (such as correlations resulting from LD). However, this procedure is computationally quite expensive, as the number of tests that have to be performed are multiplied by a factor of n.

Recently, Sul and colleagues proposed a new approach for multiple-testing correction in eQTL studies, using a multivariate normal distribution to approximate the empirical null distribution that would have been generated by permutation experiments (Sul et al., 2015). In the future, this approach may provide a way of achieving results comparable to permutation experiments at greatly reduced computational complexity.

1.8. Gene expression as a quantitative trait

Despite the additional level of resolution provided by GWAS, the complex genetic mechanisms that underlie whole-organism phenotypes have proven challenging to understand. Exactly how loci identified in a GWAS affect the phenotype often remains an open question, especially as many of them are not located inside protein-coding regions (for example Easton et al., 2007).

Mutations in regulatory regions have been shown to affect a variety of phenotypes, including traits such as the bristle number in *Drosophila* (for example Skaer and Simpson, 2000, reviewed in Wray, 2007). It thus stands to reason that many of the loci identified in GWAS do not affect the amino acid composition of a protein, but instead change the expression level of genes through changes in gene regulatory regions. This change in gene expression level could then in turn affect the phenotype. Understanding the effects of mutations on gene expression

levels is thus an important step towards bridging the gap between genotype and phenotype.

With the decreasing cost of microarrays and later RNA-seq (see Section 1.4), it has now become possible to measure the expression levels of all genes of an organism in a high-throughput manner. By considering these measurements of gene expression levels as quantitative traits, a QTL study on gene expression can be performed for each individual gene using a single assay. The loci identified in such studies are called expression QTLs (eQTLs) (Schadt et al., 2003).

The multiple testing problem is even more severe in eQTL studies than in normal GWAS, as each gene represents a separate phenotype that is being tested. Consequently, if 10,000 genes are to be tested against 1 million variants, the total number of tests performed will be $10000 \cdot 10^6 = 10^{10}$ and a Bonferroni-corrected p-value threshold would be $0.01 \div 10^{10} = 10^{-12}$. Such a low threshold would result in very low statistical power, which would mean that either very large sample sizes would be required or only very large effects could be detected.

To circumvent this problem, eQTL studies often restrict the set of variants to be tested to those in close proximity to the gene. This is based on the assumptions that most *cis* regulatory regions are located near the gene and that changes to *cis*-acting gene regulation have larger, more easily detectable effects than changes to *trans*-acting gene regulation (Petretto et al., 2006; Wittkopp and Kalay, 2011). By vastly decreasing the required number of tests in this way, such proximal eQTL studies (also sometimes inaccurately called *cis* eQTL studies, see Section 1.8.1) are much better powered to detect effects close to the gene, but with the limitation of ignoring effects further away.

The first genome-wide surveys of eQTLs were carried out in the early 2000s, based on genetic linkage analysis (Brem et al., 2002; Schadt et al., 2003). A few years later, in 2007, the first modern eQTL studies using a GWAS approach were conducted in humans (Stranger et al., 2007; Dixon et al., 2007). Since then, many more eQTL studies have been performed, both in humans (Innocenti et al., 2011; Gaffney et al., 2012; Lappalainen et al., 2013) and many other organisms (Huang et al., 2009).

Gene regulation differs between tissues, environments, developmental stages and cell types (The GTEx Consortium, 2013). For example, a SNP that causes a change to a CRM may be identified as an eQTL in a study of liver cells. However, if the transcriptional regulator that binds to this CRM is only expressed in the liver, the SNP will have no effect in another tissue. A similar scenario could

be imagined for different environments or developmental stages, where an eQTL found in a given environment or a given developmental stage may not necessarily have any effect in another.

Thus, a particular focus of eQTL studies has been the identification of eQTLs specific to various conditions, such as cell types (Dimas et al., 2009), tissues (Nica et al., 2011), temperature (Li et al., 2006) and differentiation state (Gerrits et al., 2009). The relationship between eQTLs and the developmental stage of a whole organism has only been investigated in one study so far (Francesconi and Lehner, 2014). Comparing the gene expression profiles of 200 recombinant inbred lines of *Caenorhabditis elegans*, Francesconi and Lehner tested the association of genetic markers with the expression levels of 15,855 genes. By including the estimated developmental time point of each sample as a covariate in their test, they were able to increase the number of proximal eQTLs that they could identify by 54 %. This led them to conclude that the developmental stage is an important factor in gene expression dynamics and the mapping of eQTLs. However, the samples used in this study were not staged, which meant that developmental time point of each sample had to be estimated from its gene expression profile. In addition, the mapping resolution of the study was quite limited, since it was performed on individuals from recombinant inbred lines instead of a wild population.

1.8.1. *cis*, *trans*, proximal and distal

As described in Section 1.3, regulatory effects can usually be divided into those that act in *cis* (on the same molecule of DNA) and those that act in *trans* (on a different molecule of DNA, often through an intermediate). While eQTLs located close to the gene are often assumed to act on gene expression in *cis*, a normal eQTL study cannot provide any direct evidence of this. Thus, while it has become common practice to describe eQTLs located close to their gene as *cis* eQTLs and those further away as *trans* eQTLs, this terminology is not accurate.

To identify true *cis* and *trans* effects, differences in gene expression levels between two individuals can be compared to the level of allele-specific expression observed in their F_1 hybrid, similar to a classical *cis-trans* complementation test (McManus et al., 2010; Goncalves et al., 2012). In this test, a real *cis* effect would result in differential expression between the parents and corresponding allele-specific expression in the offspring. On the other hand, a real *trans* effect would result in no allele-specific expression in the offspring.

In this thesis, I will use the terms *proximal* and *distal* to describe genetic varia-

tion close to and further away from the gene of interest, respectively. Only when there is additional evidence for the mode of action of a given variant will I refer to it as *cis* or *trans*.

1.9. The *Drosophila* Genetic Reference Panel

The *Drosophila* Genetic Reference Panel (DGRP, Mackay et al., 2012) is a collection of inbred lines sampled from a natural population of *D. melanogaster*. It consists of 205 lines that were generated from 205 individual flies collected at a farmer's market in Raleigh, North Carolina. After more than 20 generations of inbreeding, each of these 205 lines is approximately isogenic and homozygous at all variants.

Each inbred line provides a replenishable stock of genetically identical individuals on which many different kinds of experiments can be performed. At the same time, since the lines were generated from a naturally breeding population, the genomic variation between the lines is a good representation of the true spectrum of allele frequencies present in the wild. The genomes of all lines in the DGRP have been fully sequenced and genotype calls based on these sequences, including SNPs, indels and SVs, have been made available (Huang et al., 2014). Thus, the DGRP provides a unique resource to study the effects of genetic variation on many different phenotypes under many different conditions.

The DGRP is particularly well suited for the fine mapping of traits in a GWAS, as there is a very low degree of linkage disequilibrium in the population. This is reflected in the average measure of correlation between two variants, the squared Pearson correlation r^2, decaying to less than 0.2 after only 10 bp on autosomes (Graveley et al., 2011). This is much lower than for other organisms such as humans, where the r^2 decays to 0.2 only after approximately 30 kb (The International HapMap Consortium, 2007), or wild mouse populations, where the average r^2 only falls below 0.5 at a distance of 380 kb (Yalcin et al., 2010). While this high degree of recombination would be disadvantageous for classical linkage studies, it enables very fine-grained mapping of traits in a GWAS context, sometimes even to individual nucleotides. Thus, the DGRP has already powered GWAS of various phenotypes, including susceptibility to viral infection (Magwire et al., 2012), sleep (Harbison et al., 2013), pigmentation (Dembeck et al., 2015), regulation of growth (Vonesch et al., 2015) and life span (Ivanov et al., 2015).

The only eQTL study in the DGRP to date was published in 2012 (Massouras

et al., 2012). In this study, Massouras and colleagues found proximal eQTLs for 2,033 out of 7,889 tested genes at an FDR of 10 %. However, this study only considered a single developmental stage, adult flies. In addition, it used data from only 39 individuals and was thus underpowered to detect anything but the strongest effects. I will compare the results from Massouras and colleagues to my study in Chapter 5.

1.10. An eQTL study in *Drosophila melanogaster* during embryo development

In collaboration with the Furlong group at EMBL in Heidelberg, Germany, we designed a project to study the genetics of gene regulation during *Drosophila* development. Taking advantage of the unique resource provided in the DGRP we not only conducted an eQTL study at a single point in development, but also compared different stages of embryogenesis. This constitutes, to my knowledge, the first eQTL study that compares gene regulation at different stages of embryo development.

We selected three different time points for our study that represent three major stages of *Drosophila* embryogenesis:

- 2–4 h after fertilisation when the embryo transitions from the syncytium stage into the cellularised blastoderm stage and gastrulation occurs (morphological stages 5–9)
- 6–8 h after fertilisation when the germ band expands and forms visible segments (morphological stages 11–12)
- 10–12 h after fertilisation when the germ band has retracted and organ precursor cells begin to express cell-type specific markers and differentiate (morphological stages 13–15)

My collaborator Enrico Cannavò, a PhD student from the Furlong group, extracted RNA from whole-embryo samples of 80 different DGRP lines at each of the three time points. By sequencing these samples using 3′ Tag-Seq we were able to both estimate gene expression levels and also generate a comprehensive, high-resolution map of 3′ UTR cleavage sites in these stages of *Drosophila* development.

In this thesis, I will describe how I processed this 3′ Tag-Seq data into gene and 3′ transcript isoform expression levels (Chapter 2), how the transcriptome differed between the developmental stages (Chapter 3), how I detected gene-proximal eQTLs (Chapter 4), how I filtered and assessed the quality of these eQTLs (Chapter 5), what properties these eQTLs had (Chapter 6) and finally how I conducted a genome-wide search for eQTLs acting in *trans* (Chapter 7).

A paper lead by Enrico Cannavò and myself, describing parts of this project as well as additional analysis by my collaborators, has recently been submitted for peer review (Cannavo et al., 2015).

2. Processing of 3′ Tag-Seq data

3′ Tag-Seq enables the estimation of gene expression levels as well as the identification of individual 3′ transcript end sites. In this chapter, I describe how I processed the raw 3′ Tag-Seq data to locate these transcript ends and obtained per-gene and per-transcript isoform expression levels.

2.1. Introduction

To generate the data required for our eQTL study, we sequenced the RNA of whole embryos from 80 different DGRP lines at three different stages of embryo development: 2–4 h, 6–8 h and 10–12 h after fertilisation. In total, this data set comprised 254 samples, including all 240 possible combinations of the 80 DGRP lines and the three developmental stages, as well as 14 additional replicates (see Table A.1). All samples were sequenced on an Illumina HiSeq 2000 machine, multiplexing ten samples per sequencing lane. Fragments from each sample were tagged with a 6 bp barcode to allow the identification of the source sample for each sequencing read. After demultiplexing, the number of reads in the samples ranged between 8.0 million and 22 million (mean 13 million) for a total of 3.4 billion reads (Figure 2.1) with an effective read length of 43 bp.

An initial analysis showed that, after mapping to the genome, the 3′ Tag-Seq reads piled up in a peak shape, with their summit close to the 3′ end of transcripts, exactly as expected from the 3′ Tag-Seq protocol. Figure 2.2 shows the median 3′ Tag-Seq read coverage for a typical gene with a single 3′ isoform, together with the median RNA-seq coverage for comparison. A peak of reads around the 3′ of the gene is clearly visible in 3′ Tag-Seq, while reads are distributed across all exons in RNA-seq.

Standard pipelines for processing and quantifying RNA-seq data such as RSEM (Li and Dewey, 2011), Cufflinks (Trapnell et al., 2012) or HTSeq-count (Anders et al., 2015) assume that reads are distributed along the transcript body and that one molecule of RNA can yield multiple reads, depending on its length. Therefore,

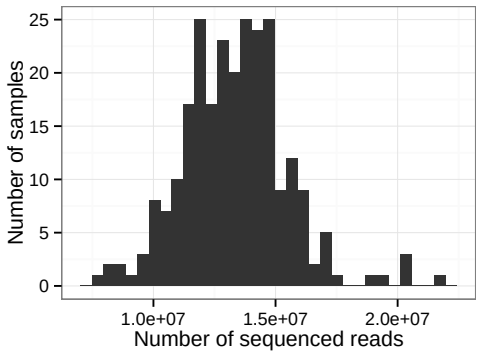


Figure 2.1.: Raw read counts in the 254 3′ Tag-Seq samples. Most samples have between 10 million and 15 million reads, with a mean of 13 million reads.

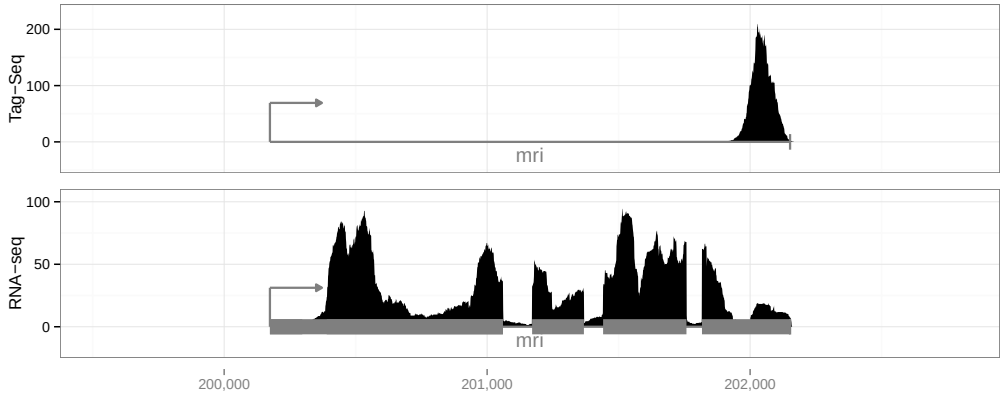


Figure 2.2.: 3′ Tag-Seq reads and RNA-seq reads for the gene *mrityu* at 10–12 h after fertilisation. Top: Median 3′ Tag-Seq read coverage across all 80 lines. Bottom: Median RNA-seq read coverage across a subset of 22 lines. Gene body shown as grey arrow, exons shown as grey boxes.

these methods were not directly applicable to the 3′ Tag-Seq data.

Estimation of expression levels of 3′ Tag-Seq data can be as simple as counting the number of reads per transcript using a reference annotation, as this number is expected to be directly proportional to the transcript abundance (Yoon and Brem, 2010). However, as I was also interested in discovering novel 3′ transcript ends this approach would not have been appropriate. Thus, I developed a custom analysis pipeline to identify 3′ transcript isoforms and quantify their expression level using 3′ Tag-Seq data, similar to the approach described in Smibert et al. (2012).

2.2. Mapping biases in eQTL studies

eQTL studies are sensitive to biases in the estimation of gene expression levels that are correlated with genomic features. For example, the estimation of gene expression levels from RNA-seq usually requires the reads to be mapped to a reference genome sequence. Reads overlapping a variant will not exactly match the reference genome if they were obtained from individuals with a non-reference genotype, and may thus fail to be mapped to the correct location. Consequently, one might observe a difference in the number of mapped reads between individuals with the reference and the non-reference genotype. Such a bias due to a difference in *mappability* may lead to the incorrect conclusion that this variant is an eQTL, while in reality it is only affecting one's ability to accurately measure gene expression levels (Degner et al., 2009).

The effect of mappability bias is exacerbated by two features in my project: Firstly, there is a large amount of indels in the *D. melanogaster* genome, which can have a larger effect on read mapping than SNPs. In the variant annotation for the DGRP (Huang et al., 2014), 657,494 indels with length greater than 4 bp are identified, which corresponds to an average of one indel approximately every 183 bp of the genome. Thus, each of the 3′ Tag-Seq peaks, which have a median width of 226 bp (see Section 2.4), can be expected to contain at least one indel. Secondly, because of the relatively small size of the 3′ Tag-Seq peaks, a single inconveniently placed variant can strongly affect the estimated expression level for the entire transcript. In normal RNA-seq, these sorts of effects would be less severe because expression levels are estimated across the whole transcript body.

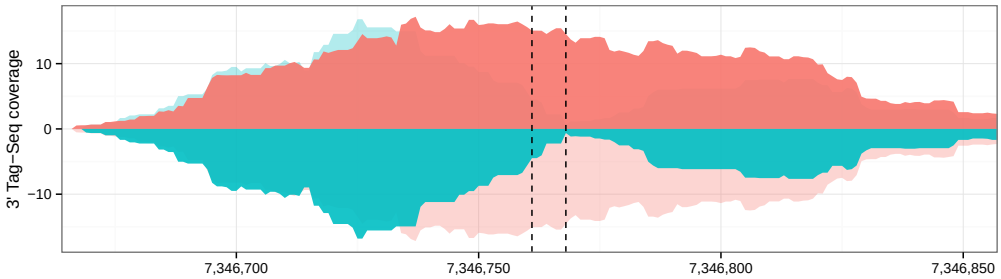


Figure 2.3.: Example of differential mappability in the 3′ Tag-Seq peak for the gene *CTCF*. 3′ Tag-Seq coverage for individuals with the reference genotype shown in red (positive values) and light red (opposite) and for individuals with the non-reference genotype shown in blue (negative values) and light blue (opposite). Black dashed lines show the location of two annotated indels in the peak region.

An example of a difference in gene expression levels associated with indels is shown in Figure 2.3. One can clearly see that the coverage for individuals with the reference genotype (red, positive) forms a normal peak shape as expected, while the coverage for the non-reference genotype (blue, negative) drops to zero in the region around the two indels (dashed vertical lines). This region of low mappability thus results in an unusual "double-peak" shape and greatly decreased gene expression level estimates for individuals with the non-reference genotype.

I employed two measures to minimise the chance of false positive eQTLs caused by mapping biases. My first step was to map the 3′ Tag-Seq reads to personalised genomes instead of the reference genome, which I will expand on in the following section. The second was to filter eQTLs by multiple criteria related to the mappability of the associated peak region. I will describe these steps in more detail in Chapter 5.

2.3. Mapping of short reads to personalised genomes

To prevent variants from affecting the mappability of the genome, I mapped the 3′ Tag-Seq reads to a custom personalised genome for each individual DGRP line. For this, I developed a tool that generates a personalised genome FASTA file for each individual listed in a variant annotation VCF file, based on a reference genome.

My tool iterates through the variant annotation at the same time as reading the reference genome and produces a custom reference sequence for each individual, containing the alternate allele at each variant where the individual had a homozygous alternate genotype. For heterozygous variants, I retained the reference sequence. This procedure is outlined in Algorithm 1.

I applied this program to the standard *D. melanogaster* reference genome (BDGP5, Adams et al., 2000) and the full binned variant annotation from the DGRP, which had been generated by the DGRP consortium based on whole-genome DNA-sequencing data (Huang et al., 2014).

The generation of personalised genomes is complicated by structural variants that result in indels, as the addition or deletion of parts of the genomic sequence will shift the coordinate systems between different lines. Because of this shift, the same gene can be located in a different location for each line, which makes comparisons between different lines, or in fact to the reference genome, challenging. Nevertheless, I did want to include indels in the personalised genomes, since they

Data: reference genome $\mathbf{r}$ (N), variant annotation $\mathbf{V}$ $(I \times N)$
Result: personalised genomes $\mathbf{G}$ $(I \times N)$
foreach *nucleotide* $n \in \langle 1 \dots N \rangle$ **do**
  foreach *individual* $i \in \langle 1 \dots I \rangle$ **do**
    switch *type of* $\mathbf{V}_{i,n}$ **do**
      case *homozygous alternate*
        $\mathbf{G}_{i,n} = \mathbf{V}_{i,n}$
      end
      case *heterozygous*
      case *homozygous reference*
        $\mathbf{G}_{i,n} = \mathbf{r}_n$
      end
    endsw
  end
end

Algorithm 1: Simplified algorithm for the generation of personalised genomes using SNP data. N is the number of nucleotides in the genome, I is the number of individuals.

had the largest chance of introducing mapping biases as described in Section 2.2.

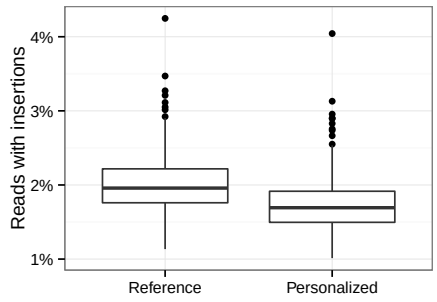
To handle this shifting of coordinates, I implemented a two-step personalisation process. First, I generated personalised genomes using only the annotated SNPs, which by definition did not result in any changes to the coordinate system. I used these "SNP-personalised" genomes to determine the location of all 3′ Tag-Seq peaks, which I will describe in Section 2.4. However, I only used these genomes to find the locations of the peaks, not to quantify their expression levels. Thus, there was no risk of introducing biases in expression levels due to indels.

Once I had determined the location of all 3′ Tag-Seq peaks in my data, I generated a second set of personalised genomes that considered both the annotated SNPs as well as annotated indels up to a size of 100 bp. When I generated these "indel-personalised" genomes, I also generated a personalised 3′ Tag-Seq peak annotation for each DGRP line, adjusting the coordinates of the original 3′ Tag-Seq peaks to shift with the insertions and deletions. I then mapped all reads to these indel-personalised genomes and quantified the expression levels in each sample using the personalised annotation, as described in Section 2.7. This approach allowed me to account for indels, while still being able to map peaks between DGRP lines, despite their shifted coordinate systems.

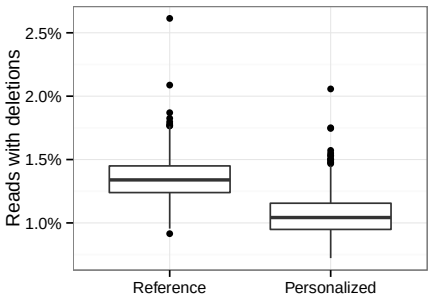
Comparing alignment statistics between mapping to the indel-personalised genomes and the reference genome, I observed that the alignments of reads to per-

sonalised genomes tended to contain fewer gaps. On average, 2.0 % of reads mapped to the reference genome contained an insertion, compared to 1.8 % of reads mapped to the personalised genome (Figure 2.4a, Student's t-test, $p = 3.71 \cdot 10^{-13}$). The same was true for deletions, with the mean percentage of reads with a deletion dropping from 1.4 % to 1.1 % (Figure 2.4b, Student's t-test, $p < 2.2 \cdot 10^{-16}$).

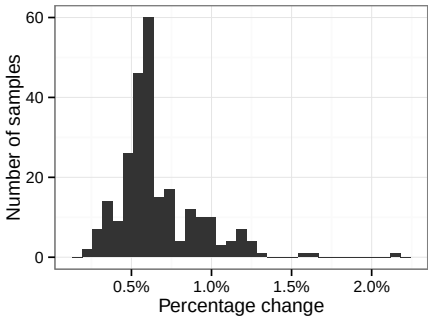
As a result, the overall number of reads that could be mapped to the genome increased by 0.66 % on average (Figure 2.4c).



(a) Insertions



(b) Deletions

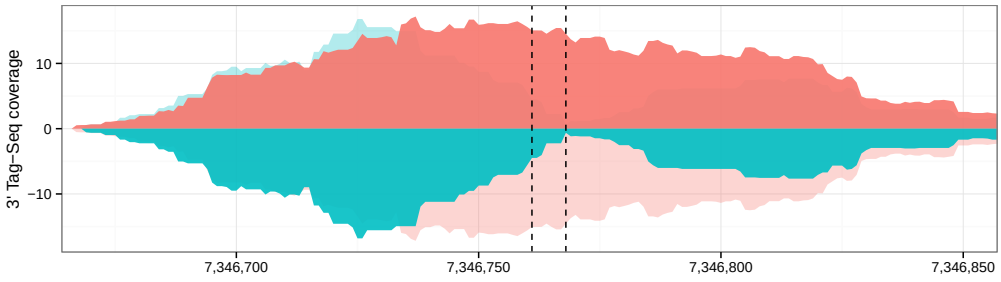


(c) Mapped reads

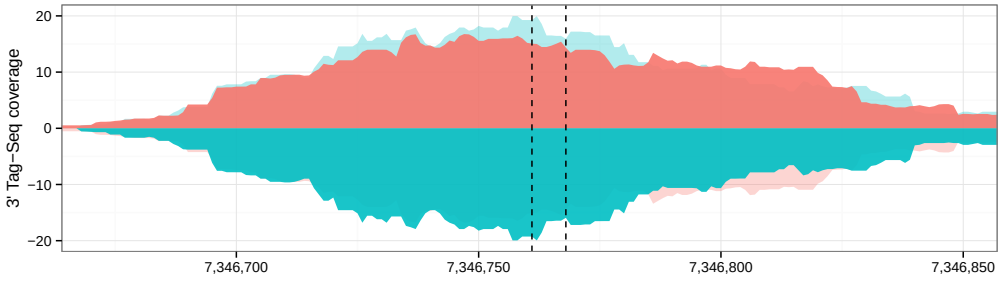
Figure 2.4.: Comparisons between reference genome and personalised genomes. (a) Percentage of read alignments that involved an insertion compared between reference and personalised genomes. (b) Percentage of read alignments that involved a deletion compared between reference and personalised genomes. (c) Increase in number of uniquely mapped reads when mapping to personalised instead of reference genomes.

An increase of 0.66 % may seem quite small, and one could argue that this might not have been worth the effort. However, this aggregate number does not do justice to the difference that the personalised genomes can make in preventing false positives caused by differential mappability. Figure 2.5 shows an updated

version of the 3′ Tag-Seq peak shown in Figure 2.3. Based on the reads mapped to the reference genome (Figure 2.5a), there seemed to be a large difference in expression level between the reference and the non-reference allele at the given variant. In fact, in an early version of the eQTL calls, I had listed this gene as having an eQTL with the insertion and the deletion inside the peak as the most strongly associated variants. However, the reads mapped to the personalised genomes (Figure 2.5b) revealed that this effect was entirely caused by a difference in mappability introduced by these variants. Once the insertions and deletions have been accounted for, the estimated expression level no longer significantly differed between the genotypes. This example shows how the genome personalisation step allowed me to account for reference mapping biases and thus allowed me to improve the quality of my expression level estimates.



(a) Reads mapped to reference genome



(b) Reads mapped to personalised genome

Figure 2.5.: Example of differential mappability of the 3′ Tag-Seq peak region for the gene *CTCF* due to indels. 3′ Tag-Seq coverage for individuals with the reference genotype shown in red (positive values) and light red (opposite) and for individuals with the non-reference genotype shown in blue (negative values) and light blue (opposite). Black dashed lines show the location of two annotated indels in the peak region.

2.4. Identification of 3′ transcript end regions from 3′ Tag-Seq poly(A) reads

The first step in determining gene expression levels was to determine the genomic location of the 3′ ends of transcripts in an annotation-agnostic way, from the 3′ Tag-Seq data only. In theory, all regions where several 3′ Tag-Seq reads map should indicate the presence of a polyadenylated transcript's 3′ end, the 3′ cleavage site. However, a known problem with 3′ Tag-Seq is that the oligo-dT primer used during the poly(A) capture and reverse transcription step can sometimes anneal to fragments containing a run of genomic poly(A) (Nam et al., 2002). Consequently, simply considering every cluster of 3′ Tag-Seq reads as evidence of a transcript end would have resulted in false positives. I thus needed to find a way to differentiate reads corresponding to real transcript ends from artefacts caused by genomic poly(A) or other experimental noise.

Since the fragmentation step yielded fragments of varying size, our 3′ Tag-Seq data contained some reads produced from very short fragments, which started with a few nucleotides from the 3′ end of a transcript but then continued with a long stretch of A nucleotides. This stretch of A nucleotides was generated from the poly(A) tail itself. I called these reads polyadenylated reads (poly(A) reads). If I could find these poly(A) reads and see where on the genome their non-poly(A) part mapped, I would be able to identify the location of the true 3′ cleavage sites. The method I developed for this purpose is similar to the one described in Smibert et al. (2012).

In order to identify poly(A) reads that truly came from polyadenylated transcript ends and not from genomic poly(A), I first mapped all reads to the SNP-personalised genomes (see Section 2.3) in their full form using the BWA short-read aligner (v0.6.2-r126, Li and Durbin, 2009). To account for sequencing errors and variants that were not considered in the genome personalisation I allowed for up to 5 mismatches and up to 10 gap extensions (parameters `-n5 -e10`) for the read mapping. I also set the option `-q20` to let BWA automatically trim reads if the sequence quality fell below 20.

Any read stemming from a genomic poly(A) tract should have corresponded to a real genomic sequence, so it should have been successfully mapped in this step. Reads covering a poly(A) tail added during mRNA processing, however, will not have had a corresponding genomic sequence and should have failed to map.

To obtain these reads, I extracted all reads that BWA failed to map to the

genome in their full form and ended in at least five A-nucleotides. I trimmed all trailing A-nucleotides off these reads and then attempted to map their remaining parts again as described above. I considered every read that could only be mapped in the trimmed form a true poly(A) read. A flow chart of this process is shown in Figure 2.6.

As expected, only a small subset of reads in each sample fulfilled these criteria (average: 0.88 %). However, since we sequenced almost 3.4 billion reads in total I was still able to obtain approximately 30 million poly(A) reads after pooling all samples.

By determining the location of the most 3′ nucleotide of each of these poly(A) reads, I could now identify the location of putative 3′ transcript ends in the *D. melanogaster* genome, down to a single nucleotide. It should be noted, however, that any A nucleotides at the 3′ end of the transcript would have been trimmed off together with the poly(A) tail, which means that I would have been unable to detect the exact 3′ end of a transcript if it ended in one or more genomic A nucleotides.

I observed 91,345 distinct putative transcript ends (cleavage sites) supported by at least 15 poly(A) reads each, which is more than six times as many as the 14,297 putative cleavage sites reported in a previous study of polyadenylation in the *D. melanogaster* genome (Smibert et al., 2012). This is not surprising given the large difference in sample sizes, but shows the additional degree of detail that this study can contribute to the genome annotation, even at a relatively high threshold of 15 reads. To simplify the downstream analysis I grouped nearby transcript ends together into clusters, considering each region with at least 15 overlapping poly(A) reads on the same strand a 3′ Tag-Seq peak region. I assume that each such peak region corresponds to a distinct poly(A) site. The 3′ end of these peaks is defined as the most 3′ cleavage site that I observed in the cluster.

Based on the expected fragment size and the observation that most non-poly(A) 3′ Tag-Seq reads mapped within 200 bp of the 3′ transcript end, I extended these peak regions upstream by 200 bp to capture all associated reads. The median summit of the 3′ Tag-Seq peaks was located 90 bp upstream of the 3′ end of the peak. When two peak regions were located within 200 bp of each other, I extended the peak only up to the 3′ end of the next upstream cluster, to make sure no overlaps occurred. In total, I identified 37,025 such 3′ Tag-Seq peaks with a median width of 226 bp. An example of the raw data used in this process is shown in Figure 2.7 and an overview over the whole process is shown in Figure 2.8.

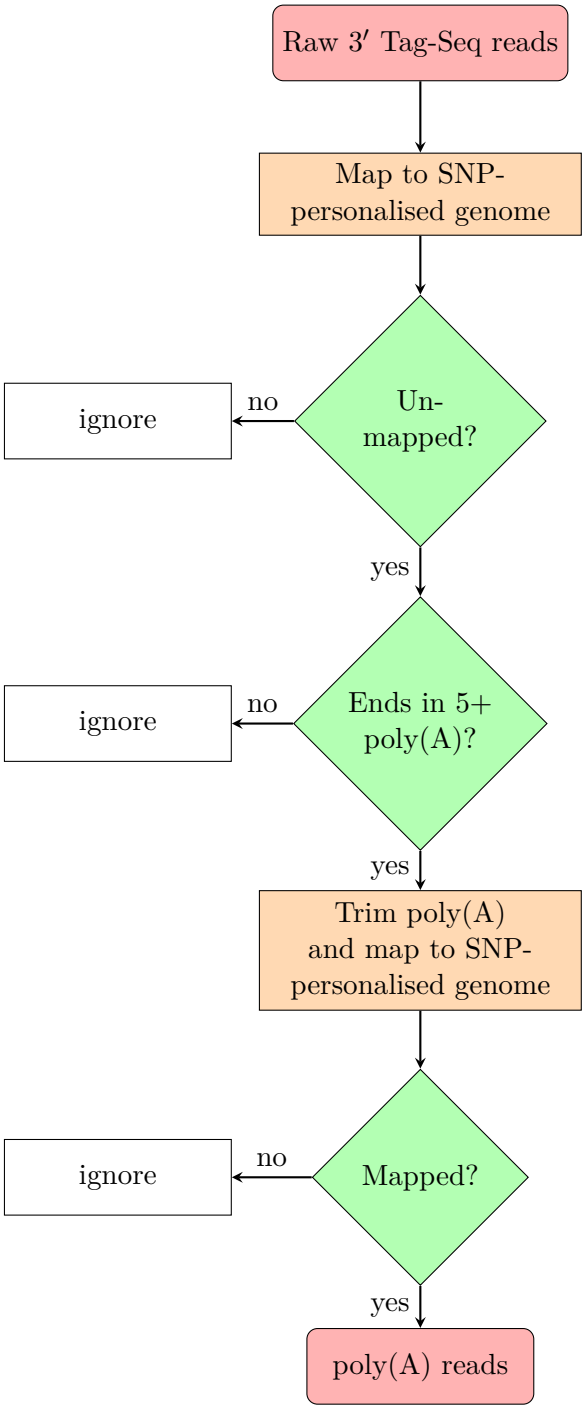


Figure 2.6.: Flow chart illustrating the processing of raw 3′ Tag-Seq reads for a single sample to the mapped poly(A) reads.

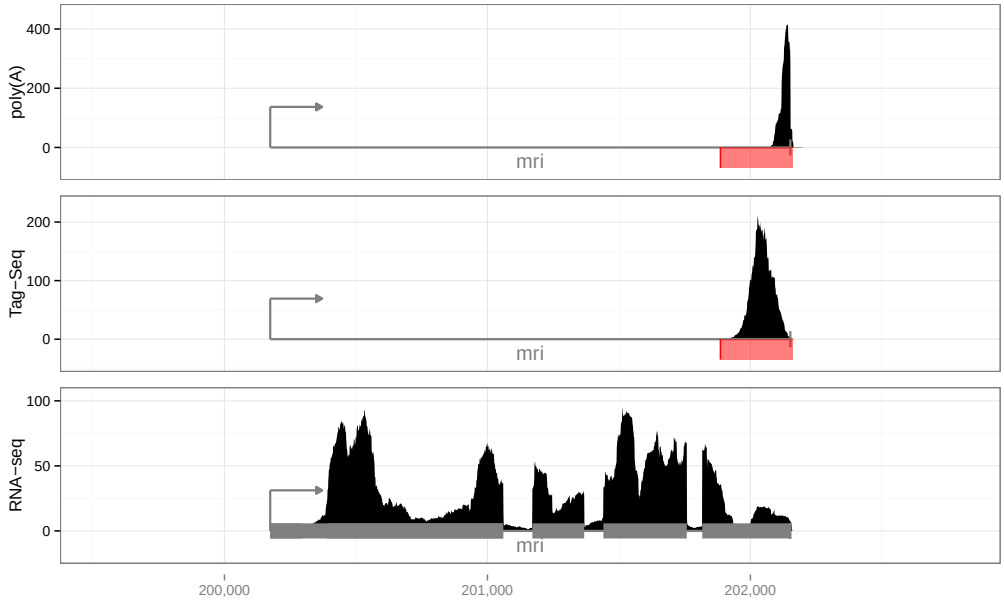


Figure 2.7.: poly(A) reads, 3′ Tag-Seq reads and RNA-seq reads for the gene *mrityu*. Top: Coverage of poly(A) reads from all samples. Middle: Median 3′ Tag-Seq read coverage at 10–12 h after fertilisation. Bottom: Median RNA-seq read coverage at 10–12 h after fertilisation. 3′ Tag-Seq peak region marked in red.

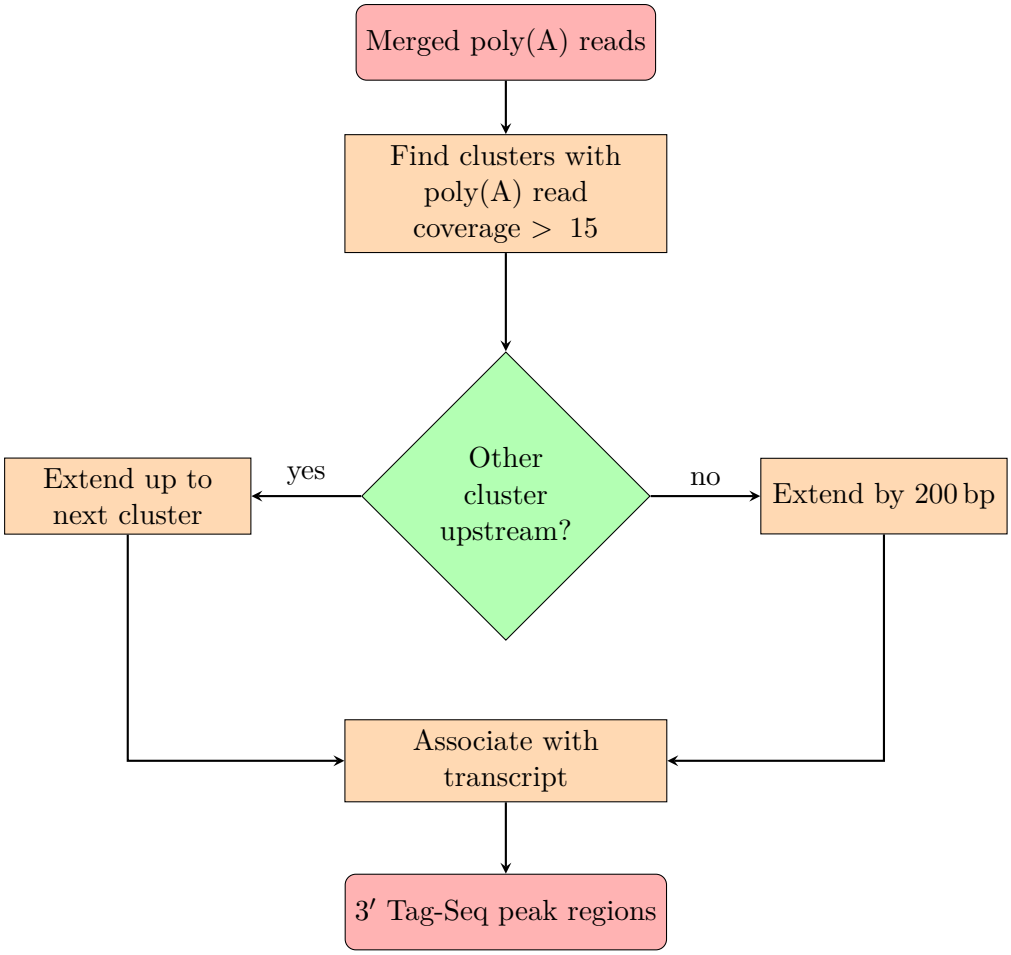


Figure 2.8.: Flow chart illustrating the identification of annotated 3′ Tag-Seq peak regions from the merged set of poly(A) reads from all samples.

2.5. Annotation of 3′ Tag-Seq peaks

Once I had identified the 3′ Tag-Seq peaks, I needed to assign them to genes. This was important both for interpreting the biological meaning of the eQTLs I would later identify as well as for estimating per-gene expression levels, which are calculated as the sum of all peak expression levels associated with that gene (see Section 2.7). For this process I considered all genes annotated on the standard chromosome arms (X, 2L, 2R, 3L, 3R, 4) in FlyBase annotation version 5.47 (The FlyBase Consortium, 2014).

Because the *D. melanogaster* genome is quite dense and the genome annotation is not always accurate, finding the correct gene for each peak was often not a straightforward task. On one hand, there were cases where the peaks I observed did not correspond to any known 3′ end of a transcript, while on the other hand sometimes a peak could have been generated by any of several different annotated transcripts. To decide which annotated transcript most likely produced each of the 3′ Tag-Seq peaks, I applied a hierarchical procedure. First I looked for the annotation most likely to result in a peak — the end of an annotated transcript on the same strand as the peak. If a peak had such an end annotated within a 500 bp window, I assigned it to the closest one. If not, I moved on to the second-most likely annotation and so on. These annotations were, in order:

1. Peak within 500 bp up-/downstream of annotated mRNA 3′ end
2. Peak within 500 bp up-/downstream of annotated ncRNA
3. Peak within 500 bp downstream of an mRNA 5′ end (TSS)
4. Peak overlapping annotated mRNA exon
5. Peak overlapping annotated mRNA intron
6. Peak within 2,000 bp downstream of annotated mRNA 3′ end (extended 3′ end)
7. Antisense annotation:
    a) Peak within 500 bp up-/downstream of annotated mRNA 3′ end
    b) Peak within 500 bp downstream of an mRNA 5′ end (TSS)
    c) Peak overlapping annotated mRNA exon
    d) Peak overlapping annotated mRNA intron

If there were multiple possible annotations of the same rank and with equal distance to the peak, I selected one of them at random. However, the identifiers of all possible genes were provided as an additional annotation, which I used in some of the downstream analysis. Any peak that did not have any transcript annotated nearby was classified as "unannotated".

Of the 37,025 poly(A) sites I identified, 20,939 (57 %) could be assigned to a mRNA 3′ end, with the remaining 43 % split between the other categories (Figure 2.9). For 6 % of peaks I could not find any suitable annotation.

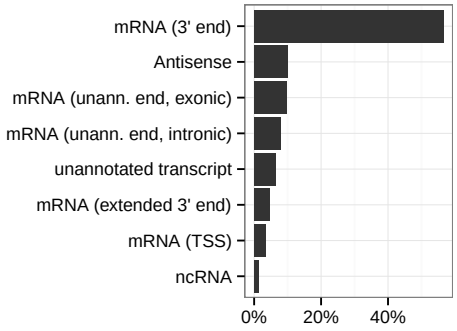
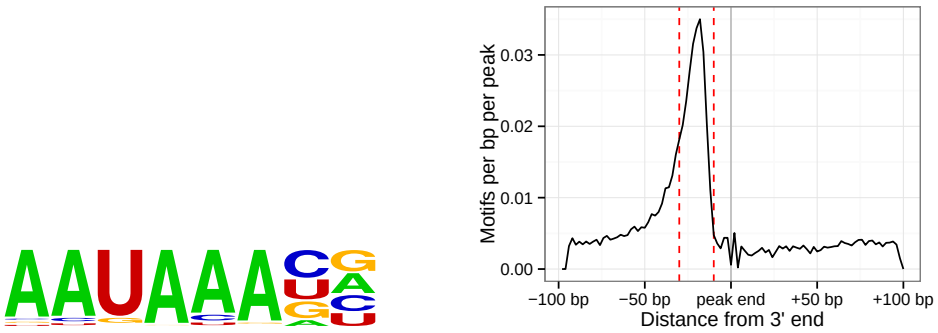


Figure 2.9.: Transcript annotation of 3′ Tag-Seq peaks.

2.6. Properties of 3′ Tag-Seq peaks

To confirm that the 3′ Tag-Seq peaks were representative of real transcript ends, I performed *de novo* motif discovery in a stranded region of 100 bp up- and downstream of the 3′ end of every peak. I used the software package Homer (Heinz et al., 2010) to identify motifs of length 6 bp or 8 bp and found 20 motifs that were significantly enriched ($p < 1 \cdot 10^{-10}$) around peaks. The most significantly enriched signal ($p = 1 \cdot 10^{-537}$) matched the canonical polyadenylation signal AAUAAA (Figure 2.10) and was present within the 200 bp window for 47 % of the peaks. The mean location of this motif was 15 bp upstream of the 3′ end, which is in line with prior work that has reported the location of the polyadenylation signal to be approximately 10–30 bp upstream of the 3′ cleavage site (see Section 1.3.2).

Most genes had a single peak associated with them, but some had up to 75 (Figure 2.11). This is similar to the data reported by Smibert et al. (2012), who had observed a similar long-tailed distribution of polyadenylation sites per gene.



(a) Sequence logo **(b)** Relative location of motif

Figure 2.10.: AAUAAA motif found through a *de novo* motif search in a 200 bp window around 3′ Tag-Seq peak ends. Red dotted lines indicate 30 bp and 10 bp upstream of the peak's 3′ end.

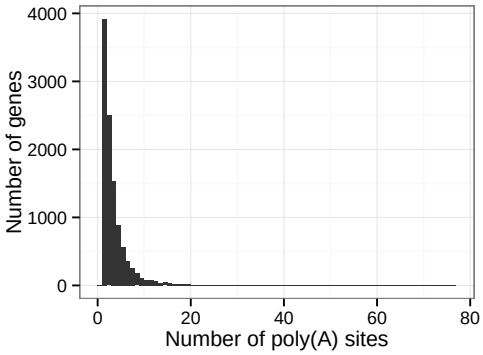


Figure 2.11.: Histogram of the observed numbers of 3′ Tag-Seq peaks per gene.

To determine if genes of a certain category were more likely to have many 3′ Tag-Seq peaks than others, I performed a gene ontology (GO, The Gene Ontology Consortium, 2000) enrichment analysis. However, to conduct this analysis in an unbiased way, it was important to account for the fact that genes tended to have more peaks the more strongly expressed they were (Figure 2.12).

This bias was most likely caused by the threshold of 15 poly(A) reads required for the identification of a 3′ Tag-Seq peak, which meant that secondary peaks of more lowly expressed genes may not have passed this threshold while secondary peaks of highly expressed genes did. To account for this, I stratified the genes into ten quantiles based on the number of 3′ Tag-Seq reads that had been assigned to them. I removed the first and the last quantile to remove outliers and then performed a GO enrichment analysis separately on each of the remaining eight

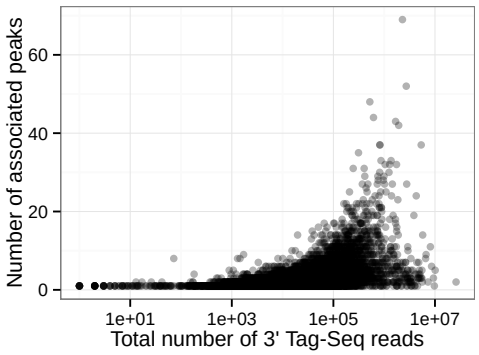


Figure 2.12.: Number of 3′ Tag-Seq peaks associated with each gene against the total number of reads assigned to the gene across all samples.

quantiles. For each quantile, I classified a gene as having many peaks if it had more peaks than 80 % of all genes in the quantile. Then, using the bioconductor package topGO (Alexa and Rahnenfuhrer, 2010), I performed a one-tailed Fisher's exact test to determine whether membership in any GO category was associated with having many peaks. I tested all GO categories from the "biological process" ontology with a minimum size of 100 annotated genes.

The GO terms "signaling" and "single organism signaling" were consistently enriched in genes with many peaks, with a p-value of less than 0.05 after Bonferroni correction in 6 out of 8 (75 %) of the quantiles. Furthermore, the GO terms "cell communication", "regulation of biological process" and "biological regulation" were enriched in 5 out of 8 (63 %) quantiles using the same threshold. This indicates that genes involved in signalling, cell communication and regulation are particularly likely to use multiple different 3′ ends regardless of expression level, suggesting that APA may play a major role for these genes.

2.7. Quantification of expression levels

Having identified the 3′ Tag-Seq peak regions, I now needed to estimate the expression level of each peak in each sample. For this, I mapped the original reads from each sample to the genomes again, this time using the fully indel-personalised version (see Section 2.3). Overall, 92 % of successfully mapped 3′ Tag-Seq reads fell into a peak region, reflecting the good sensitivity of the peaks I had identified.

I calculated the height of each peak for each sample based on the maximum number of overlapping reads and used this value as my measure of expression

level. I then summed the expression levels of all sense peaks associated with each gene, to obtain a set of total gene expression levels. While slightly different from other approaches that simply considered the total number of reads to estimate expression (Yoon and Brem, 2010), this approach proved to be a reliable measure of expression levels (see Section 2.8).

To make these expression levels comparable between samples, I finally scaled each of them by a size factor associated with each sample. Instead of using the total sum of mapped reads as in FPKM, I estimated this size factor as the 90^{th} percentile of expression levels to avoid potential biases due to outliers. Once I had divided the expression levels of each sample by this size factor, I obtained the final, scaled expression levels. Figure 2.13 shows a flow chart of this process.

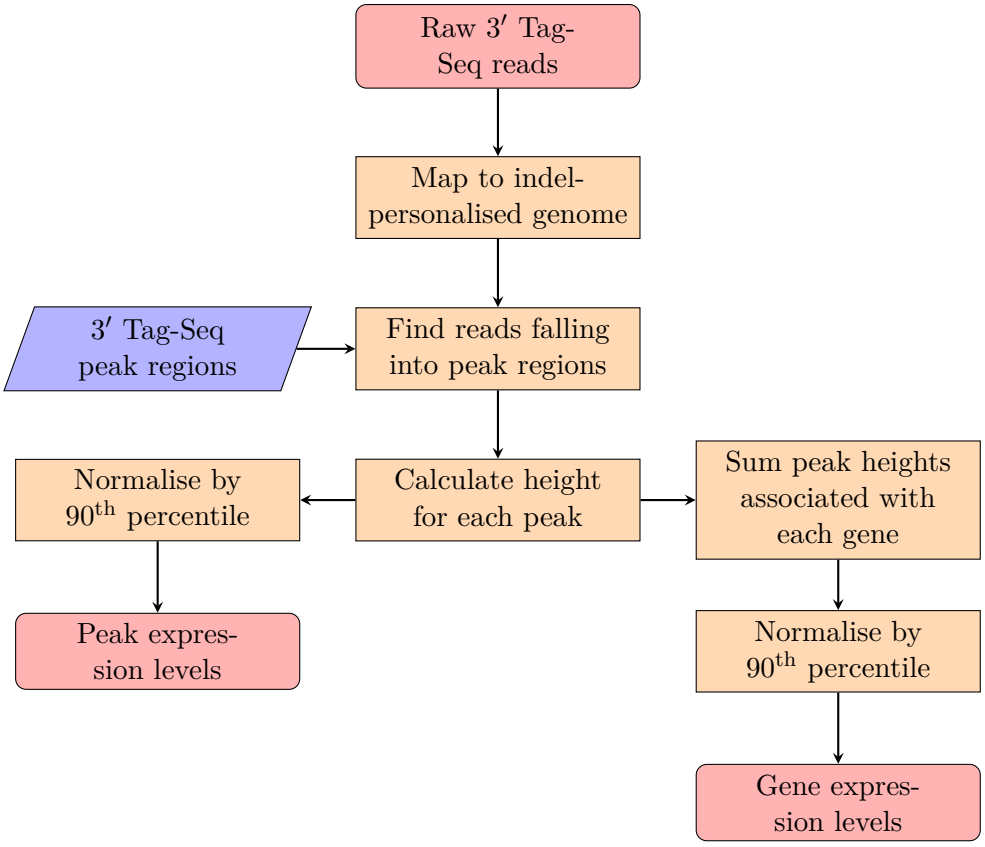


Figure 2.13.: Flow chart illustrating the quantification of 3′ Tag-Seq peak and gene expression levels for a single sample.

2.8. Comparison of 3′ Tag-Seq to standard RNA-seq

The standard RNA-seq protocol is a widely used and reproducible method of estimating gene expression levels (Marioni et al., 2008). In order to determine how reliable the 3′ Tag-Seq gene expression level estimates were, I thus compared them to estimates obtained from standard RNA-seq.

For this, my collaborator Enrico Cannavò prepared a set of 22 Illumina poly(A)$^+$ strand-specific RNA-seq libraries of samples collected at 10–12 h post-fertilisation. The libraries were prepared from the same samples of RNA that had already been used for the 3′ Tag-Seq protocol, allowing me to treat them as technical replicates to isolate protocol-specific effects. The fragments from 10 of the samples were sequenced from a single end while the ones from the other 12 samples were sequenced from both ends, yielding 10 single-end and 12 paired-end data sets. The two groups of samples were multiplexed to 12 samples per lane and then sequenced in two separate runs of a Illumina NextSeq 500 sequencing machine. In total, we sequenced 360 million fragments for the single-end samples and 344 million fragments for the paired-end samples.

I demultiplexed and quality-filtered these data sets using reaper and tally (Davis et al., 2013). After demultiplexing, the number of fragments per sample ranged between 6.6 million and 38 million (mean 27 million).

Using STAR (Dobin et al., 2013) I then mapped the reads to the *D. melanogaster* reference genome and estimated gene expression levels, measured in transcripts per million (TPM), using Cufflinks v2.2.1 (Trapnell et al., 2012).

A scatter-plot of the expression levels estimated by RNA-seq and 3′ Tag-Seq for a sample of line 517 at 10–12 h is shown in Figure 2.14. Only genes with at least some expression observed with both methods are shown, I will discuss genes for which I had observed no expression at all with one of the methods below.

Spearman's correlation coefficient between the RNA-seq and the 3′ Tag-Seq expression levels for this sample, considering all genes that showed at least some level of expression with both methods, was 0.88 ($p < 2.2 \cdot 10^{-16}$). In the 22 samples I tested, this correlation coefficient ranged between 0.80 and 0.92 (mean 0.87).

To get an overview over the total degree of correlation between the methods, I also compared the 5 % trimmed mean expression levels observed in the 22 RNA-seq libraries to the 5 % trimmed mean expression levels observed in the corresponding 3′ Tag-Seq libraries. A scatter plot of the mean 3′ Tag-Seq expression level against

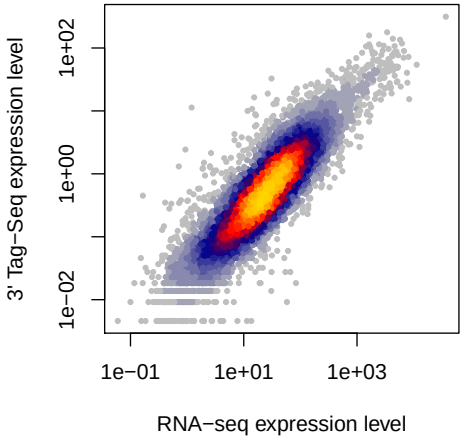


Figure 2.14.: Scatterplot of gene expression levels estimated with standard RNA-seq and 3′ Tag-Seq for line 517 at 10–12 h. Spearman's $\rho = 0.88$.

the mean RNA-seq expression level of each annotated gene is shown in Figure 2.15. Spearman's correlation coefficient for this data was $\rho = 0.90$ ($p < 2.2 \cdot 10^{-16}$).

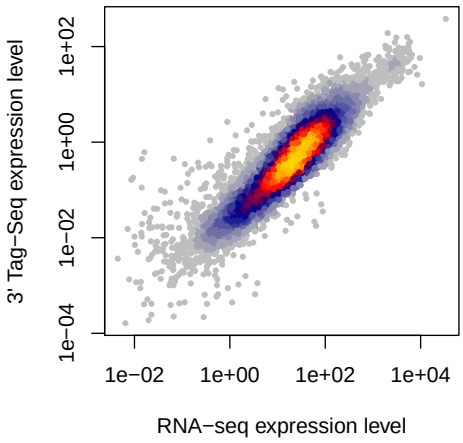


Figure 2.15.: Scatterplot of mean gene expression levels estimated with standard RNA-seq and 3′ Tag-Seq across 22 different lines at 10–12 h. Spearman's $\rho = 0.90$.

There were, however, 1,953 genes expressed according to RNA-seq for which I had estimated an expression level of zero in the 3′ Tag-Seq data. Most of the strongly expressed genes that I did not observe in 3′ Tag-Seq were cases in which I had in fact observed a 3′ Tag-Seq peak, but where I had been unable to assign

this peak to a unique gene. This was the case for example for the gene *sala*, which in the Flybase annotation has the exact same genomic coordinates as the gene *CG43355*. As I ignored all peaks that could not be assigned to a unique gene, it was impossible for me to observe gene expression in such a case of ambiguous gene annotation. However, there were only 124 genes for which this was the case.

The remaining genes that I had only observed with RNA-seq were much more lowly expressed than the genes I had observed with both methods. The median RNA-seq expression level of the genes without ambiguous annotation but only seen in RNA-seq was 0.16 TPM, while the median of all other genes was 20.15 TPM (Figure 2.16, Wilcoxon test, $p < 2.2 \cdot 10^{-16}$).

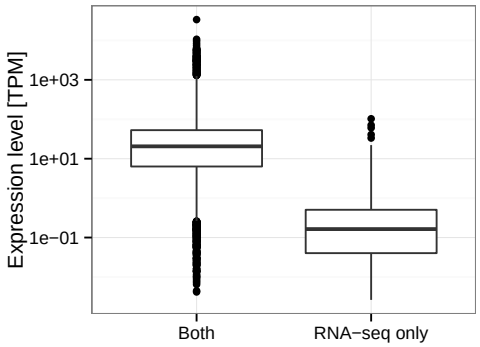


Figure 2.16.: Comparison of RNA-seq expression level of genes observed with both RNA-seq and 3′ Tag-Seq and those only observed with RNA-seq.

The fact that I only saw expression of such lowly expressed genes in the RNA-seq data was not surprising, given the fact that I had obtained more than twice as many reads per sample in the RNA-seq analysis as in the 3′ Tag-Seq analysis (3′ Tag-Seq reads per sample = 13 million, RNA-seq reads per sample = 27 million). Thus, I was able to detect the levels even of very weakly expressed genes, which I missed in the 3′ Tag-Seq data. A greater sequencing depth for the 3′ Tag-Seq data would have allowed me to test these genes for eQTLs as well, but would have also essentially doubled the cost of the sequencing for this project. Since this would have meant that I would have been able to sequence fewer different lines, which would have reduced my power to detect eQTLs, I believe that the omission of some lowly expressed genes was an acceptable compromise.

In summary, the 3′ Tag-Seq gene expression levels generally appeared to be a good estimate of the true gene expression levels, as confirmed by RNA-seq. However, there was some disagreement between the methods, which may have

been caused by experimental variation, problems such as the assignment of a $3'$ Tag-Seq peak to an incorrect gene or sequencing biases inherent to one of the protocols. I will analyse the quality and properties of these expression levels further in Chapter 3.

3. Analysis and normalisation of gene expression levels

In this chapter I describe how I compared the gene expression levels I obtained to a reference time course to remove mis-staged and low-quality samples from my data set. I then explore differences in gene expression levels between the three developmental stages and discuss how I normalised the expression levels to make them suitable for eQTL testing (Chapter 4).

3.1. Introduction

The variation in gene expression levels between different individuals of *Drosophila melanogaster* at the same point in development is generally low (Rifkin et al., 2005). Within each of the three developmental stages, I should thus have obtained similar gene expression level profiles for all samples. Hence, any sample with a gene expression profile that was very divergent from the expected profile was likely to be an outlier and not a reflection of genetic differences. Identifying and removing these cases was particularly important for my eQTL study, as outliers would decrease my statistical power to detect true genetic effects.

In addition, the samples collected for my study were staged solely by the time since fertilisation of the egg (2–4 h, 6–8 h and 10–12 h), not by the phenotype of the embryos. Thus, I needed to confirm that all DGRP lines did not only follow the same developmental programme, but also developed at the same speed, with each nominal time point reflecting the same stage of development in all samples.

3.2. The developmental transcriptome of *D. melanogaster*

In 2011 the modENCODE consortium performed poly(A)$^+$ RNA-seq on embryos from the *D. melanogaster* reference line at various developmental time points, including at each 2-hour interval between 0 and 24 hours post-fertilisation (Graveley et al., 2011). The gene expression levels measured in this study show that

different genes follow different trajectories of gene expression during development, with some of them being constitutively expressed and others switching on and off for different developmental stages. These expression levels thus form a unique gene expression profile at each developmental stage.

As a first step of my analysis I calculated the pairwise Spearman correlation between the embryonic gene expression levels (measured in FPKM) provided by the modENCODE consortium, which is shown as a heatmap in Figure 3.1.

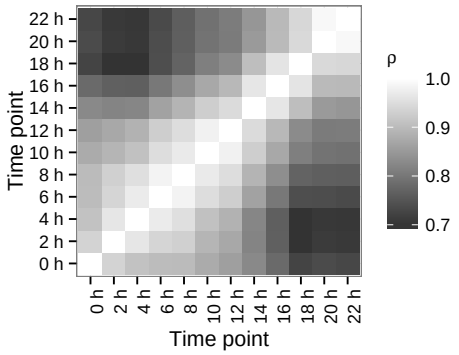


Figure 3.1.: Heatmap of pairwise Spearman's correlation coefficients between all time points in the modENCODE embryonal time course.

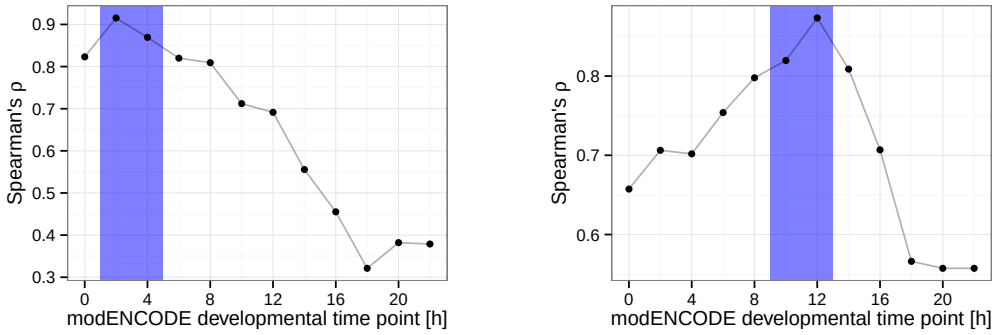
This heatmap shows not only that there are clear differences in gene expression levels between the developmental time points, but also that two time points are more similar to each other the less time has passed between them.

3.3. Staging by comparison to a developmental time course

I thus reasoned that, as long as the 3′ Tag-Seq gene expression levels correlated reasonably well with the modENCODE gene expression levels, I should be able to determine which developmental stage a sample came from, simply by finding the modENCODE reference time point it was most similar to.

To test whether this method worked, I applied it to two representative samples from our 3′ Tag-Seq gene expression data set, one collected at 2–4 h post-fertilisation and one collected at 10–12 h post-fertilisation. For each sample, I calculated the ranked correlation coefficient (Spearman's ρ) between the 3′ Tag-Seq expression levels and the modENCODE expression levels at each of the reference time points. I only considered genes that were observed in at least one 3′ Tag-Seq

sample and at least one modENCODE sample. I then plotted the correlation coefficient ρ against the reference time points. The two examples are shown in Figure 3.2.



(a) 2–4 h **(b)** 10–12 h

Figure 3.2.: Spearman's ρ between gene expression levels of 3′ Tag-Seq samples for line 461 with each embryonic modENCODE time point. modENCODE time points that match the collection time point highlighted in blue.

The correlation estimates followed a peak shape, indicating that the sample correlated best with a single time point, and then became more different the further I moved away from it. This is concordant with the assumption that expression levels would change systematically over time. Most importantly, the location of the peak was different for the two samples, and in fact corresponded to one of the two reference time points closest to their collection time point. This experiment thus showed that it was feasible to determine the developmental time point of a sample based solely on its gene expression levels.

I now extended this approach to all 3′ Tag-Seq samples to answer two questions: How well do the expression levels of each sample correlate with the reference expression levels in general and from which developmental time point is each sample. The correlation estimate ρ for the original set of samples and all embryonic modENCODE time points is shown in Figure 3.3. For each sample, I ranked the reference time points by their ρ, under the assumption that the most strongly correlated time point was the most likely true time point of the sample. If the true time point matched the collection time point, I considered the sample validated. Additionally I also considered samples validated if the second-best correlation matched the reference time point, in order to account for minor variations in the timing of both my study as well as modENCODE.

Considering all of the embryonic modENCODE time points, the highest cor-

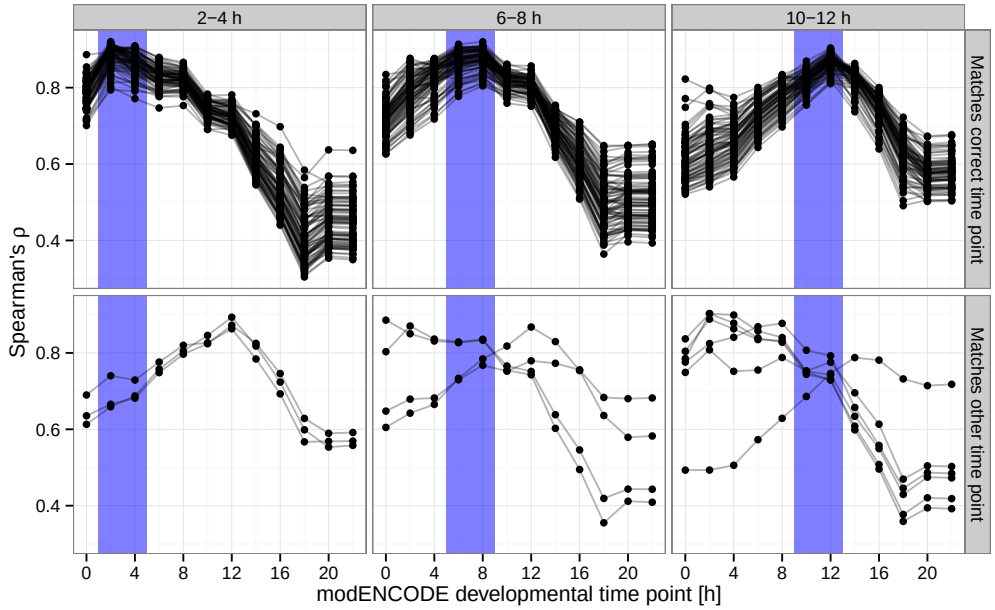


Figure 3.3.: Spearman's ρ between gene expression levels of each 3′ Tag-Seq sample with each embryonic modENCODE time point. modENCODE time points that match the collection time point highlighted in blue. Vertically split by the time point the samples were collected at, horizontally split by whether one of the two highest ρs was observed for a matching time point or not.

relation coefficient ρ ranged between 0.78 and 0.92 (mean 0.88) for the samples. This showed that most of our samples were of good quality and represented biological states similar to the ones from the reference time course. For almost all of the samples the time point estimated from the modENCODE correlation was concordant, with the collection time point matching the highest ρ for 238 samples (93 %) and the second-highest ρ for another 5 samples (2 %). This showed that the "developmental clock" seemed to run at approximately the same speed for almost all of DGRP lines we analysed, with none of them developing noticeably faster or slower than the reference.

However, some samples appeared to have a good correlation to one of the modENCODE time points, but not to the one as which they were labelled. For example, the three mismatching samples from the 2–4 h time point were clearly not correlated with the 2–4 h reference time point, but would have perfectly matched the 10–12 h time point. In a separate analysis, Jacob Degner, a postdoctoral fellow from the Furlong group at EMBL-Heidelberg, also observed that these samples seemed to have been assigned to the wrong DGRP line. This was based on a com-

parison of SNP genotypes observed in the 3′ Tag-Seq data to known genotypes from the DGRP variant annotation.

After some further investigation, it emerged that these were actually mislabelled samples, which I could correct and assign back to their correct line and time point. The new correlation coefficients after recovering the swapped samples are shown in Figure 3.4. For 242 out of 254 samples (95 %) the correct reference time point now was the time point with the highest ρ, and for an additional 5 samples (2 %) it was the second-highest. For the other seven samples, I assumed that there might have been an error during sample preparation, sequencing or processing and thus removed them from all further analysis. The maximum ρ of the remaining 247 samples ranged between 0.79 and 0.91 (mean 0.87). Thus, using this method of developmental staging by gene expression, I was not only able to identify and remove outliers from my study, but could also identify and correct sample swaps.

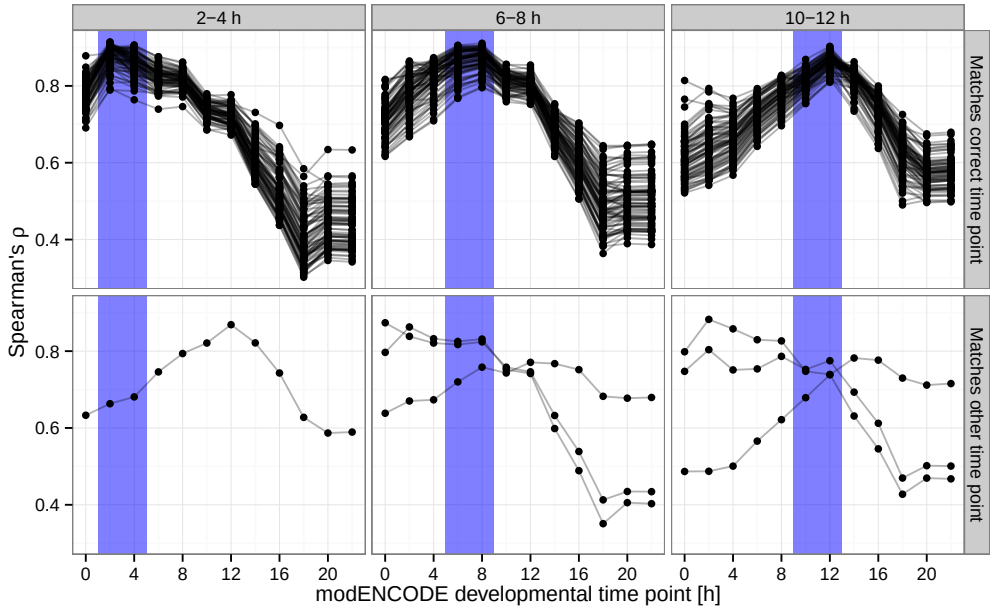


Figure 3.4.: Spearman's ρ between gene expression levels of each 3′ Tag-Seq sample with each embryonic modENCODE time point. modENCODE time points that match the collection time point highlighted in blue. Vertically split by the time point the samples were collected at, horizontally split by whether one of the two highest ρs was observed for a matching time point or not.

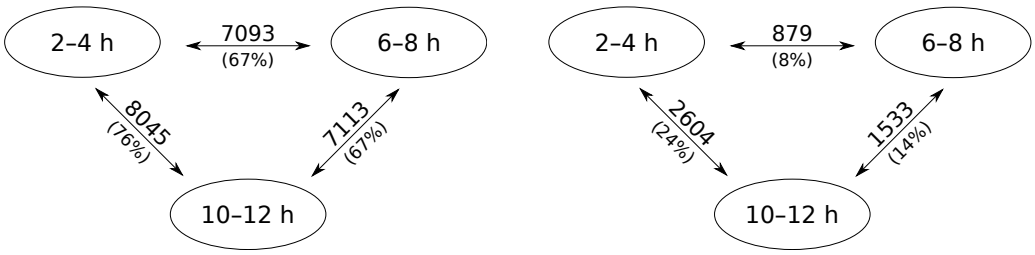
Nevertheless, I could still observe some slight differences between the samples in both the location of the best correlation and the strength of the correlation. There are various factors that may have caused this, ranging from small differences

in the developmental speed of different DGRP lines to technical reasons such as sequencing problems or variation in exactly at what point in the two-hour time range the sample was collected. However, since I could only compare our samples to the 2 h bins from modENCODE I could not resolve this structure any further using this method. I will discuss a possible extension of this approach for a more fine-grained staging of samples in Chapter 8.

3.4. Differential gene expression between developmental stages

To further investigate the differences that I could observe between developmental stages, I conducted a differential gene expression study between all samples from 2–4 h, 6–8 h and 10–12 h after fertilisation. Since each line from the DGRP constituted an independent biological sample, I had approximately 80 replicates for the gene expression levels at each developmental stage, allowing me to estimate both the mean expression level and the biological variation with very high confidence.

Using DESeq2 (Love et al., 2014) on the matrix of raw summit heights (which are comparable to the read counts usually used with DESeq2) I was able to detect a large amount of differential gene expression between the three developmental stages (Figure 3.5).



(a) all **(b)** $|\log_2(\text{fold-change})| > 1$

Figure 3.5.: Genes differentially expressed between the three developmental stages at an FDR threshold of 1 %. (a) All differentially expressed genes. (b) Genes with an absolute $\log_2$(fold-change) of at least 1.

Between 67 % and 76 % of tested genes were found to be differentially expressed between the stages at an FDR threshold of 1 %, with 8 % to 24 % different by more than a factor of two. The difference was smaller between adjacent time points (2–4 h and 6–8 h, 6–8 h and 10–12 h) than for 2–4 h and 10–12 h. These results

are in line with the results from the modENCODE project, which had already shown that the transcriptome undergoes massive, but systematic, changes during embryonic development.

Genes differentially expressed with at least a two-fold change between 2–4 h and 6–8 h were enriched in 35 different GO terms in the "biological process" ontology and depleted in another 24 (Two-tailed Fisher's exact test, Bonferroni-corrected p-value < 0.05). Many of the enriched terms, shown in Table 3.1 were related to the regulation of transcription, and RNA biosynthetic processes, while the depleted terms were related to organelle organisation, metabolic processes, translation, and the cell cycle. Between 6–8 h and 10–12 h there were no enriched GO terms, but a total of 73 depleted GO terms, many of them the same as between 2–4 h and 6–8 h (Table 3.2). The reason for this consistent depletion may be that genes from these categories are important throughout the life cycle of the embryo, and are thus constitutively expressed. The large amount of enrichment between 2–4 h and 6–8 h may be reflective of the major changes that occur in the organism between 2–4 h and later stages of development, including cellularisation and the maternal-zygotic transition (see Section 1.2.1).

This large number of differences in gene expression was reflective of the scale of change that can occur within 2 h of *Drosophila* embryo development. In addition, it also showed the large statistical power that this data set provides, making it possible to detect even low fold-changes in gene expression with high confidence.

3.4.1. Expression levels of stage-specific genes

The list of genes differentially expressed between the developmental stages included many classic developmental genes. In this section, I will show a few examples.

twist (Figure 3.6) is a helix-loop-helix transcription factor that plays a major role during gastrulation, when it is important for the formation of the mesoderm (Thisse et al., 1988; Murre et al., 1989). In accordance with this function, it was highly expressed at 2–4 h but much more lowly expressed at 6–8 h and 10–12 h. The difference in expression level between 2–4 h and 6–8 h was highly significant according to DESeq2 (BH adjusted p-value $= 4.65 \cdot 10^{-187}$), while the expression level at 10–12 h was too low to quantify the difference.

pebbled (Figure 3.7), also known as *hindsight*, is a zinc finger transcription factor involved in the retraction of the germ band, which begins during the 6–8 h time point and completes at the beginning of the 10–12 h time point (Yip et al., 1997).

Term	Exp.	Obs.	Dir.	$-\log_{10}(p)$
regulation of transcription, DNA-dep...	64.94	121	↑ enr.	9.08
regulation of RNA biosynthetic process	64.94	121	↑ enr.	9.08
regulation of transcription from RNA p...	33.35	75	↑ enr.	8.44
regulation of macromolecule biosynthetic...	70.99	124	↑ enr.	7.38
regulation of cellular macromolecule bio...	70.99	124	↑ enr.	7.38
transcription, DNA-dependent	70.62	122	↑ enr.	6.85
RNA biosynthetic process	70.81	122	↑ enr.	6.82
organic cyclic compound biosynthetic pr...	87.30	142	↑ enr.	6.60
regulation of RNA metabolic process	71.09	122	↑ enr.	6.60
regulation of cellular biosynthetic pr...	73.69	124	↑ enr.	6.30
organelle organization	94.94	42	↓ depl.	8.03
cellular protein metabolic process	106.49	55	↓ depl.	6.32
RNA processing	35.78	7	↓ depl.	6.22
mRNA metabolic process	28.60	6	↓ depl.	4.07
translation	32.24	8	↓ depl.	4.06
protein metabolic process	137.14	89	↓ depl.	3.60
mRNA processing	26.93	6	↓ depl.	3.44
cell cycle phase	45.09	18	↓ depl.	3.17
mitotic cell cycle	44.63	18	↓ depl.	2.89
cell cycle process	52.64	24	↓ depl.	2.79

Table 3.1.: GO terms in the "biological process" ontology significantly enriched (enr.) or depleted (depl.) in genes that changed by at least two-fold between 2–4 h and 6–8 h. p, p-value after adjustment for multiple testing. Only top 10 out of 35 enriched terms shown. Only top 10 out of 24 depleted terms shown. Exp., expected number of genes. Obs., observed number of genes. Dir., direction.

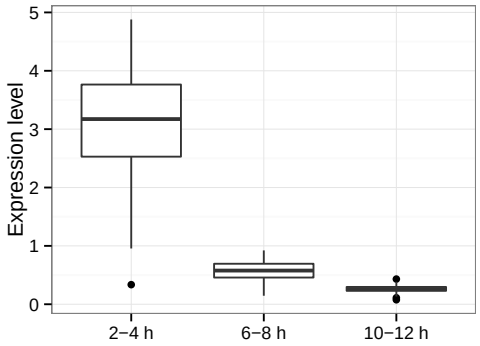


Figure 3.6.: 3′ Tag-Seq expression level of *twist* across 80 different lines from the DGRP.

Accordingly, *pebbled* expression was strongly increased between 2–4 h and 6–8 h (BH adjusted p-value $= 3.53 \cdot 10^{-69}$) and then decreased again between 6–8 h and

Term	Exp.	Obs.	Dir.	$-\log_{10}(p)$
gene expression	227.94	106	↓ depl.	22.82
cellular macromolecule metabolic process	346.21	210	↓ depl.	20.40
cellular metabolic process	455.73	318	↓ depl.	17.82
RNA processing	55.56	5	↓ depl.	16.77
cellular protein metabolic process	168.73	83	↓ depl.	13.52
cellular process	760.23	649	↓ depl.	13.17
RNA metabolic process	176.31	93	↓ depl.	12.00
mRNA metabolic process	44.33	5	↓ depl.	11.80
mRNA processing	41.71	5	↓ depl.	10.52
macromolecule metabolic process	408.48	303	↓ depl.	10.08

Table 3.2.: GO terms in the "biological process" ontology significantly enriched (enr.) or depleted (depl.) in genes that changed by at least two-fold between 6–8 h and 10–12 h. p, p-value after adjustment for multiple testing. Only top 10 out of 73 depleted terms shown. Exp., expected number of genes. Obs., observed number of genes. Dir., direction.

10–12 h (BH adjusted p-value $= 2.95 \cdot 10^{-18}$).

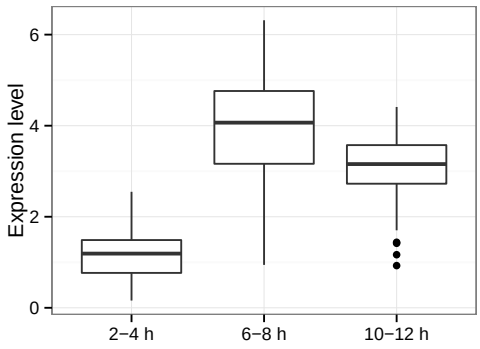


Figure 3.7.: 3′ Tag-Seq expression level of *pebbled* across 80 different lines from the DGRP.

The gene *serpentine* produces a luminal chitin binding protein, which is important for tracheal development (Luschnig et al., 2006). *serpentine* has previously been shown by *in situ* hybridisation to be expressed in tracheal cells starting from morphological stage 12 (7 h 20 min after fertilisation) and continuing until the end of embryogenesis (Wang et al., 2006). The pattern that I observed was in agreement with this, with the expression of *serpentine* being very low at 2–4 h and 6–8 h and high at 10–12 h (BH adjusted p-value $= 1.56 \cdot 10^{-124}$ between 6–8 h and 10–12 h).

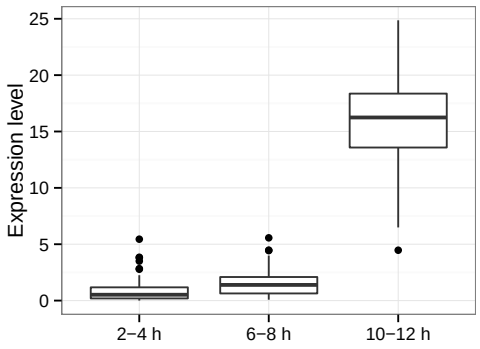


Figure 3.8.: 3′ Tag-Seq expression level of *serpentine* across 80 different lines from the DGRP.

3.5. Normalisation of 3′ Tag-Seq data for eQTL discovery

For all further analyses, I only considered the first replicate that passed the staging filter for each unique combination of line and developmental stage, resulting in a final set of 235 samples including 76 lines for which I had data at all three developmental stages. These gene expression levels now had to be normalised in a way that made them suitable for an eQTL study.

In short-read sequencing there are various potential confounding factors that can cause systematic biases in the estimated gene expression levels (Leek et al., 2010). This can result in the expression levels of samples being closely correlated with each other, for example because they were sequenced in the same sequencing lane (batch effects) or because they are closely related genetically (population structure). Large-scale correlations between samples are thus usually reflective of confounding factors as opposed to genetic effects. In this section, I will describe how I normalised the gene expression levels to remove any potential outliers and then removed confounding effects using the software package PEER.

In order to estimate the level of structure in the raw (only library-size adjusted) expression level matrix, I calculated the pairwise Spearman correlation between each pair of samples. I then applied the `hclust` function in R (R Core Team, 2013) to obtain a hierarchical clustering of the samples based on the euclidean distance between their correlation coefficient vectors.

In addition, I performed a principal component analysis (PCA) on the expression level matrix. PCA is used to transform a set of many, possibly correlated variables (such as gene expression levels) into a set of uncorrelated (orthogonal)

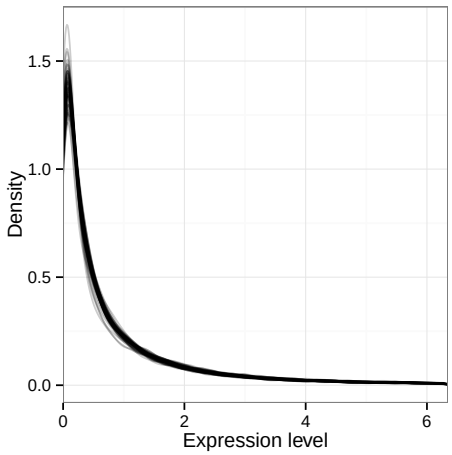
variables. These uncorrelated variables are then called the principal components (PCs). The first principal component (PC1) always explains the largest amount of variance in the original set, the second (PC2) explains the second-largest amount of variance in the original set and so on. If there is large-scale structure in the data, the first few components will explain the vast majority of the variance, while for data with more fine-grained structure each of them will only explain a little. Thus, we can get an idea of how structured a set of observations is by looking at the variance explained by its first few principal components. In addition, by projecting each sample onto the space defined by PC1 and PC2, we can get a good overview over which samples are similar to each other. PCA is sometimes used to correct for batch effects in sequencing data, however I will only be using it as a diagnostic plot.

The diagnostic plots generated with these methods for the raw expression levels of genes at 10–12 h are shown in Figure 3.9. In order to illustrate the difference between the scaled and unscaled expression levels, I did not centre and scale the expression level matrix for this first PCA. While the strength of the first component was most likely caused by this lack of centring, which can inflate the size of the first component in PCA, the size of the other components as well as the large amount of correlation between the samples showed that there was a substantial amount of large-scale structure in the data.

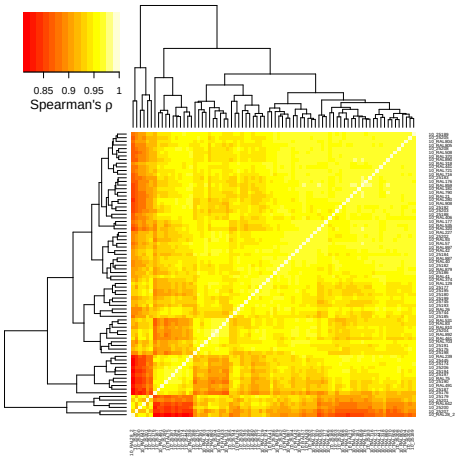
I thus applied multiple normalisation steps to our expression levels, similar to those that had been shown to increase the number of discovered QTLs by Degner et al. (2012). I applied these normalisations separately to each developmental time point and to each feature type (genes and 3′ Tag-Seq peaks) using custom R scripts. I will describe them here for the example of genes at 10–12 h and show diagnostic plots similar to Figure 3.9, using the following notation:

Let $\mathbf{Y}^{\text{method}}$ be a $N \times D$ matrix of gene expression levels at 10–12 h, where "method" is the normalisation method, N is the number of individuals and D is the number of genes/peaks. I will denote operations that are applied to each row vector i separately as $\mathbf{Y}_{i,\dots}$ and operations applied to each column vector j as $\mathbf{Y}_{\dots,j}$.

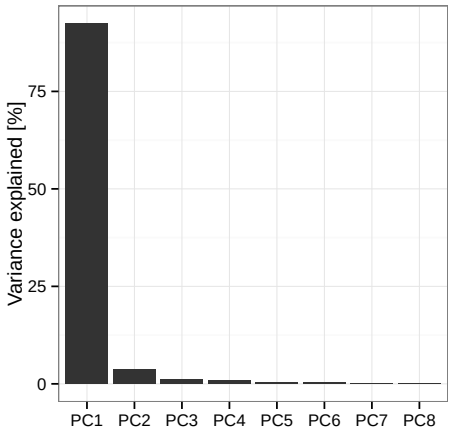
First, I removed all genes for which the median expression level (among all samples) was zero, indicating that this gene was not expressed at all in more than half of the samples. I centred and scaled the expression levels associated with each remaining gene by subtracting the mean and dividing by the standard deviation of the gene's expression level. The resulting z-score describes how much above or



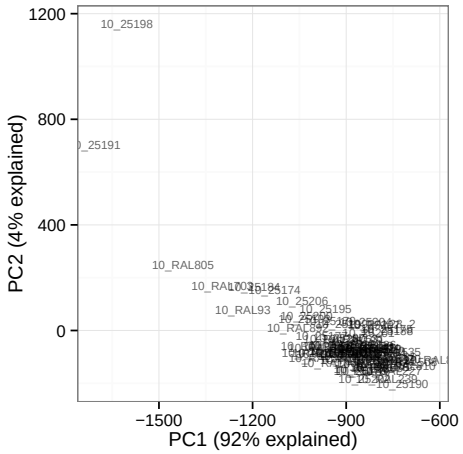
(a) Expression levels per sample, with top and bottom 5^{th} percentile trimmed



(b) Pairwise correlation between samples



(c) Variance explained by first eight PCs



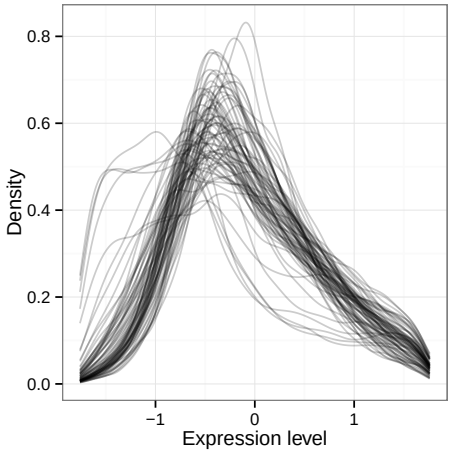
(d) Samples projected onto first two PCs

Figure 3.9.: Diagnostic plots for raw, library-size adjusted gene expression level matrix of 78 samples from the 10–12 h time point. Correlation between the samples is high (mean $\rho = 0.94$) and the first principal component explains 92 % of the variance.

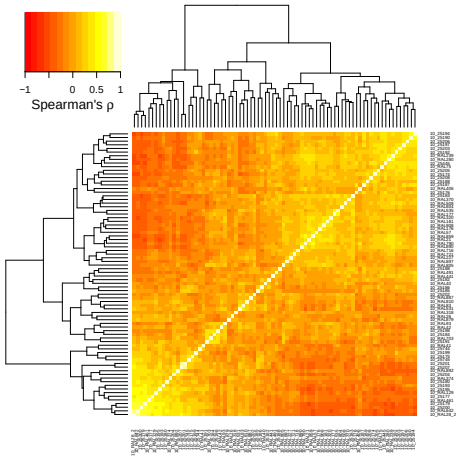
below the mean expression level each individual's expression level was, normalised by the observed variation in the gene's expression levels overall. This is shown in the equation below, where $\text{mean}(x)$ calculates the mean and $\text{sd}(x)$ calculates the standard deviation of a vector x.

$$\mathbf{Y}^{\text{z-scores}}_{\ldots,j} = (\mathbf{Y}^{\text{raw}}_{\ldots,j} - \text{mean}(\mathbf{Y}^{\text{raw}}_{\ldots,j}))/\text{sd}(\mathbf{Y}^{\text{raw}}_{\ldots,j}) \tag{3.1}$$

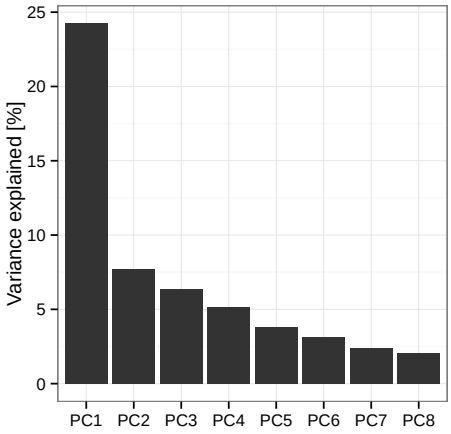
The diagnostic plots for the centred and scaled expression level matrix $\mathbf{Y}^{\text{z-scores}}$ are shown in Figure 3.10.



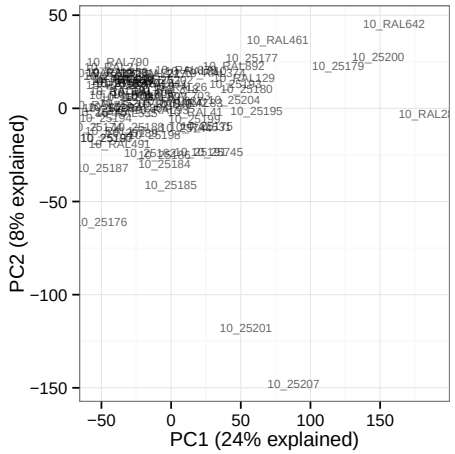
(a) Expression levels per sample, with top and bottom 5^{th} percentile trimmed



(b) Pairwise correlation between samples



(c) Variance explained by first eight PCs



(d) Samples projected onto first two PCs

Figure 3.10.: Diagnostic plots for centred and scaled (z-score) gene expression level matrix of 78 samples from the 10–12 h time point. Most samples are no longer correlated and the first principal component explains 24 % of the variance, indicating that there is less large-scale structure in the data. Expression levels are now roughly symmetric around 0, but there are some outliers.

Since the DGRP lines used in this study were all viable and did not show any large phenotypic differences (see Section 1.9), I further assumed that each individual should show average expression levels for most genes, with only a few of them being over- or underexpressed due to genetic differences. To make sure that

all individuals conformed to this assumption, I quantile-normalised the z-scores by transforming each individual's z-scores into a standard normal distribution. This is illustrated below, where qqnorm(x) is a function that takes a vector x and returns a vector z with those values transformed into a normal distribution. Figure 3.11 shows the diagnostic plots for this new matrix $\mathbf{Y}^{\text{qq}}$.

$$\mathbf{Y}^{\text{qq}}_{i,\ldots} = \text{qqnorm}(\mathbf{Y}^{\text{z-scores}}_{i,\ldots}) \tag{3.2}$$

I applied these normalisation steps to both peak and gene expression levels, and to samples from each developmental stage separately. Consequently, an individual with its expression level for a gene consistently below the mean at each developmental stage had similar normalised values in all three stages. This became important when I ran the multi-stage model to call eQTLs (see Section 4.5) since it made it feasible to observe a common effect across all three time points even if the mean expression level of the gene changed between the stages.

3.5.1. Correcting for batch effects and population structure using PEER

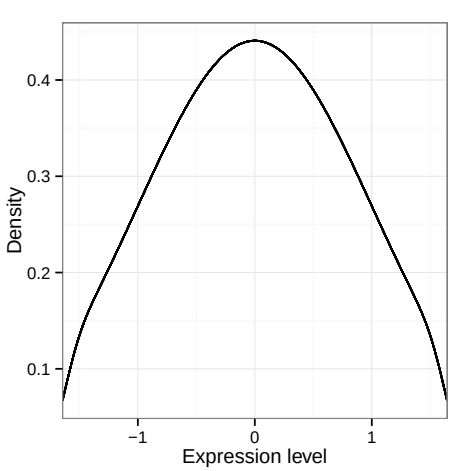
The final step in the normalisation procedure was to apply the PEER algorithm (Stegle et al., 2012) to the quantile-normalised z-scores in order to remove any remaining covariance structure from the data.

PEER uses factor analysis methods to infer hidden effects from a matrix of gene expression levels, which can then be corrected for. It assumes that there are a relatively small number of hidden factors that affect gene expression levels for many genes and individuals. Examples of such hidden effects in my study could be batch affects associated with the library preparation and sequencing, but also effects of population structure within the DGRP. Conceptually, PEER is similar to PCA, but it does not require orthogonality of factors and automatically determines the minimal number of factors necessary to model the structure, which prevents over-fitting.

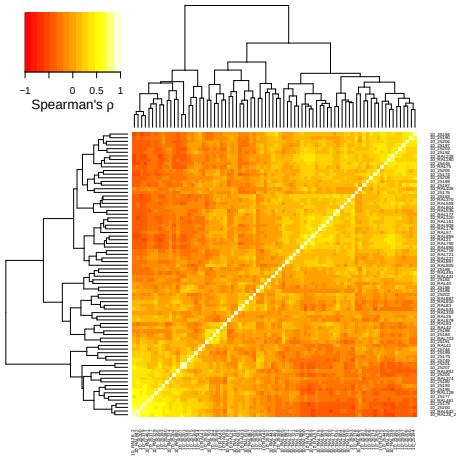
PEER models the observed gene expression level matrix $\mathbf{Y}$ (where $\mathbf{Y} = \mathbf{Y}^{\text{qq}}$) as follows:

$$\mathbf{Y} = \mathbf{XU} + \mathbf{\Psi} \tag{3.3}$$

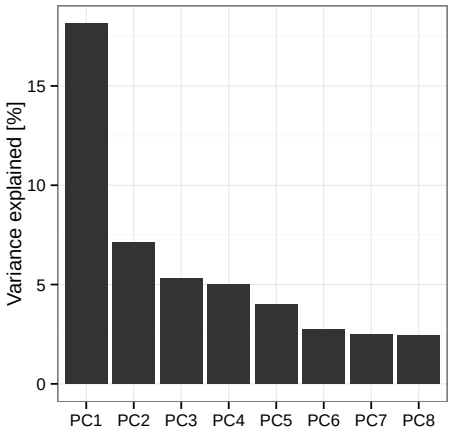
Where $\mathbf{X}$ is a $N \times k$ matrix of k hidden factors, $\mathbf{U}$ is a $k \times D$ matrix of weights for each of the factors on each gene and $\mathbf{\Psi}$ is the noise term ($N \times D$), which is



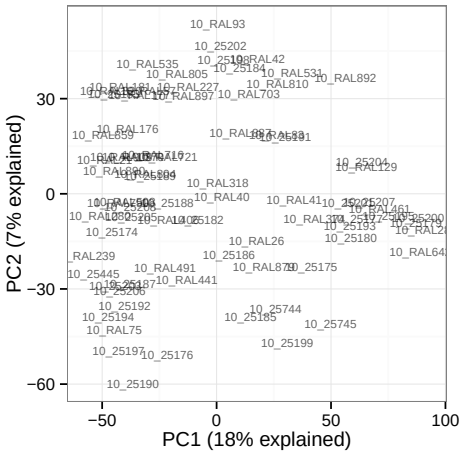
(a) Expression levels per sample, with top and bottom 5^{th} percentile trimmed



(b) Pairwise correlation between samples



(c) Variance explained by first eight PCs

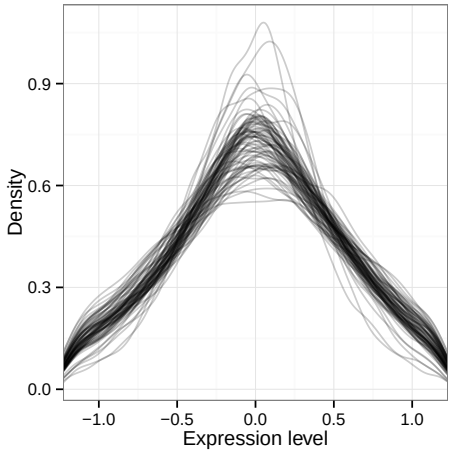


(d) Samples projected onto first two PCs

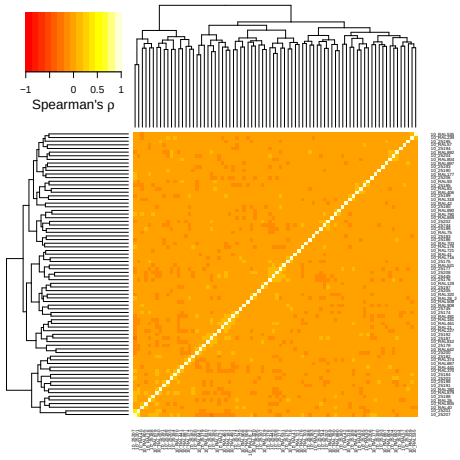
Figure 3.11.: Diagnostic plots for quantile-normalised expression level matrix of 78 samples from the 10–12 h time point. The first principal component now only explains 18 % of the variance, indicating even less large-scale structure in the data. Expression levels all follow exactly the same normal distribution, so they are symmetric.

assumed to follow a Gaussian distribution. PEER also allows for the specification of an additional matrix of covariates that can be included in the model (such as gender or age in human studies), however I did not supply any covariates in this case. The parameters of this model, $\mathbf{X}$ and $\mathbf{U}$ are initially set to values obtained via PCA and then iteratively refined using variational Bayesian learning (Stegle et al., 2010).

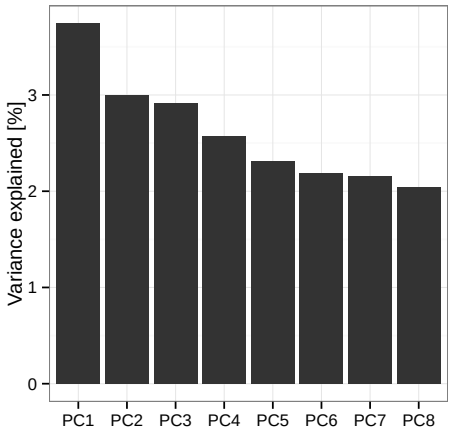
After applying PEER to my data, I obtained a matrix containing the residual gene expression levels after subtracting the effect of the hidden factors. The diagnostic plots for these PEER-normalised expression levels are shown in Figure 3.12. It is worth contrasting the diagnostic plots seen in Figure 3.9, where blocks of correlated samples could be observed, with the data shown in this figure.



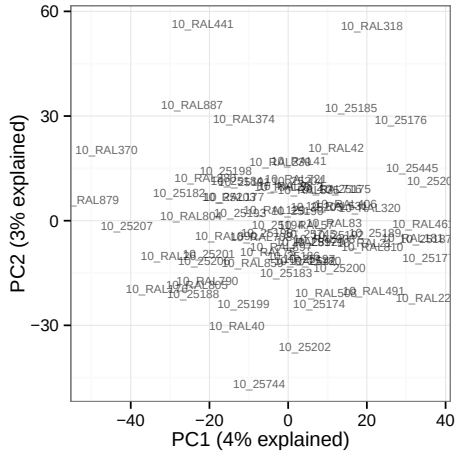
(a) Expression levels per sample, with top and bottom 5$^{\text{th}}$ percentile trimmed



(b) Pairwise correlation between samples



(c) Variance explained by first eight PCs



(d) Samples projected onto first two PCs

Figure 3.12.: Diagnostic plots for PEER-adjusted gene expression level matrix of 78 samples from the 10–12 h time point. As expected, there is almost no remaining correlation between the samples and the first principal component now explains only 4 % of the variance. The expression level distributions per sample vary slightly, but are still largely symmetric around zero.

With this final, fully normalised set of expression levels I could now go on to call eQTLs, described in Chapter 4.

4. Gene-proximal calling of eQTLs

In this chapter I describe how I used the normalised gene expression levels I estimated in Chapter 3 to find eQTLs in a proximal window around genes and 3′ Tag-Seq peaks. In the next chapter I will describe how I further filtered these eQTLs to decrease the false positive rate and remove possible artefacts introduced by the sequencing protocol, followed by an analysis of the properties of proximal eQTLs in Chapter 6. Finally, I extend this approach to test for eQTLs genome-wide in Chapter 7.

4.1. Introduction

Linear mixed models (see Section 1.7) have been shown to be a powerful tool for modelling gene expression data (Listgarten et al., 2010) and have already successfully been applied in genetic association studies to find eQTLs (Bennett et al., 2010; Lippert et al., 2014; Tung et al., 2015). The mixed model framework makes it possible to test the effect of a given variant (a fixed effect) on gene expression, while also accounting for confounders such as the population structure using random effects. An implementation of this framework is LIMIX (Lippert et al., 2014), which has been developed by the Stegle group at EMBL-EBI and allows for the rapid testing of a wide variety of univariate and multivariate linear mixed models.

In the following sections, I will describe how I used LIMIX to model the normalised expression levels as a sum of genetic and non-genetic effects and call eQTLs in the *Drosophila* Genetic Reference Panel.

The genomes of all lines in the DGRP have been fully sequenced and SNPs, indels and SVs have been annotated based on these full genome sequences (see Section 1.9). Due to the high quality of the indels and SVs in this data set (Huang et al., 2014), I not only considered all biallelic SNPs, but also all biallelic indels and SVs for my study. I defined the gene-proximal area as a ±50 kb window around each gene, starting 50 kb upstream of the 5′ end and ending 50 kb downstream of

the 3′ end. A window of this size would include almost all known *cis*-regulatory modules associated with a known gene (1,878/1,894, 99 %), as annotated in the Redfly database of *D. melanogaster* regulatory elements (Gallo et al., 2011), which suggests that most eQTLs would be located in this region. On average, this proximal region contained 1,580 annotated variants per gene, ranging from 103 to 5,601.

For all of the following analyses I only considered the 76 lines for which at least one replicate passed quality-control at each of the three developmental stages (see Table A.1), for a total for $3 \cdot 76 = 228$ samples. Additionally, I ignored any genes or peaks located on heterochromatic parts of chromosomes, as variants had not been called in these regions.

4.2. Variance decomposition

A useful first step towards finding eQTLs is to determine how much of the variance of each gene could be explained by genetic or non-genetic effects, regardless of the exact location of the associated variant.

To perform this analysis, I used the multi-trait variance decomposition module from LIMIX, modelling the expression levels of each gene as a sum of random effects: the effect of the developmental stage (Dev. Stage, σ^2_{Dev}), the effect of the local genetic relatedness based on variation in a 50 kb window (proximal, $\sigma^2_{\text{A,proximal}}$), the effect of global genetic relatedness based on variation further away (distal, $\sigma^2_{\text{A,distal}}$), as well as effects accounting for the interaction between the developmental stage and both types of genetic relatedness (G./Dev., $\sigma^2_{\text{Dev}\times\text{A,proximal}}$ and $\sigma^2_{\text{Dev}\times\text{A,distal}}$). I did not consider any non-additive genetic or epistatic effects, as they would have greatly increased the complexity of the model. I assumed that any remaining variation was either experimental or biological noise.

Since the data normalisation procedure I described in Section 3.5 was designed to make expression levels comparable between developmental stages, I would not have been able to estimate the developmental component if I had used this data set. Consequently, I used the raw, library-size adjusted, expression levels instead, which I transformed (quantile-normalised) to be normally distributed. After discussions with the authors of LIMIX I also only considered genes that were strongly expressed in all three stages (more than 90 % of samples above the 20^{th} percentile) for this analysis, in order to exclude genes that are only expressed in one of the developmental stages from inflating the estimates of the mean genetic/development

interaction effects.

This left me with a set of 6,529 genes for the variance decomposition analysis. From this data, I calculated the relative contribution of each component to the total variance of each gene and estimated the total narrow sense heritability (heritability through additive genetic effects) as the fraction of variance explained by any of the genetic effects:

$$h^2 = \sigma^2_{\text{A,proximal}} + \sigma^2_{\text{A,distal}} + \sigma^2_{\text{Dev}\times\text{A,proximal}} + \sigma^2_{\text{Dev}\times\text{A,distal}} \tag{4.1}$$

A histogram of the total narrow sense heritability is shown in Figure 4.1.

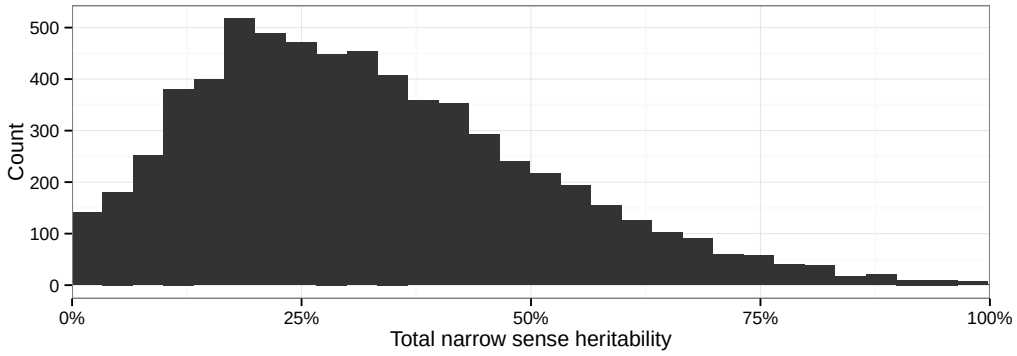


Figure 4.1.: Histogram of total narrow sense heritability h^2 for 6,529 genes, estimated as the fraction of variance explained by additive genetic effects.

The median heritability of gene expression levels was 29.8 % and 1,130 of the 6,529 genes I analysed showed a heritability larger than 50 %. I grouped the genes into bins based on the heritability of their expression levels and then plotted the mean size of each variance component for each group. The result is shown in Figure 4.2.

For 2,451 genes (38 %), more than half of their variance could be attributed to the developmental stage alone. This confirmed the large difference in gene expression levels between the three developmental stages, which I had already observed in the differential expression analysis (Section 3.4).

In addition, this plot gave me some hints about the overall structure of genetic effects that I might be able to observe in the eQTLs. First, the distal relatedness matrix appeared to explain more of the variance in gene expression levels than the proximal relatedness matrix. This suggested that there is a large number of genetic effects acting on gene expression levels in *trans*. This is in line with earlier

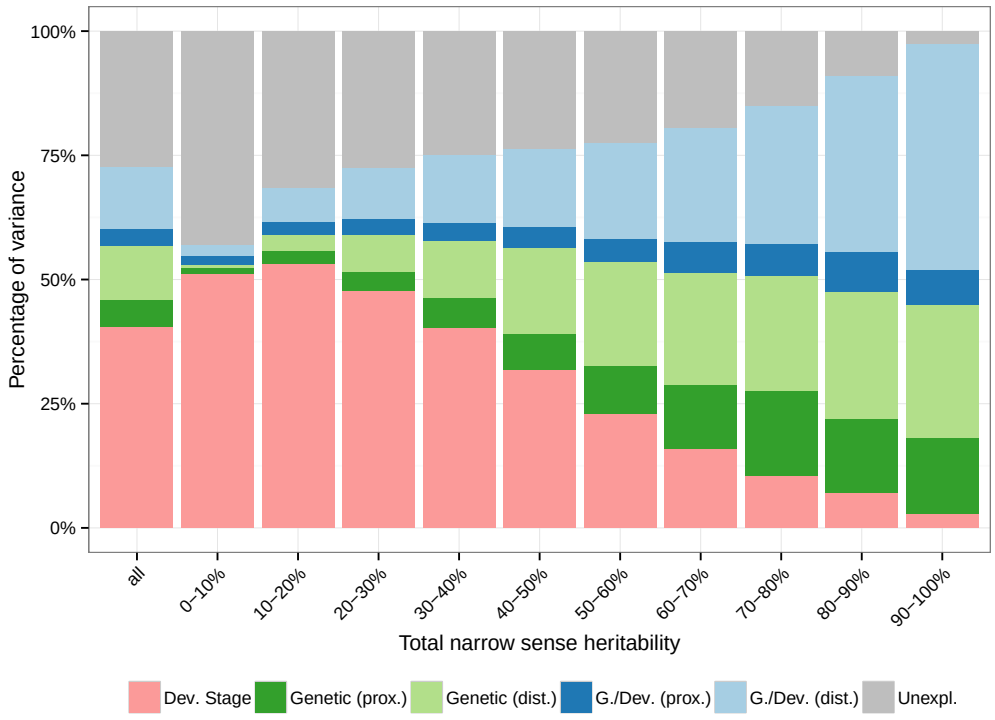


Figure 4.2.: Mean percentage of variance explained by each of the six variance components in groups of genes, binned by their total narrow sense heritability.

work in *Drosophila*, which has shown a large amount of gene regulation occurring in both *cis* and *trans*, with 66 % of genes showing evidence of *trans* regulation and 51 % of genes showing evidence of *cis* regulation (McManus et al., 2010).

The difference between distal and proximal effects was even more pronounced in the genetic/developmental stage interaction components, where the distal relatedness seemed to play a much larger role than the proximal relatedness. This suggested that developmental stage-specific regulation may largely be occurring in *trans*, while *cis* effects are usually common amongst all developmental stages. Nevertheless, I found 35 genes for which the proximal genetic/developmental stage component explained more than 25 % of the variance, which were likely to manifest also as stage-specific eQTLs. In addition, for 323 (4.95 %) of genes the pure proximal genetic component explained more than 25 % of the variance, suggesting that I would likely be able to find a common eQTL for them.

These results confirmed that there was a genetic component to the expression level of many genes. In addition, I could also see that the developmental stage

did indeed play a big role in the expression levels of the genes and that there were genetic factors whose effects were dependent on the developmental stage.

4.3. Power calculation

Before calling eQTLs I first had to answer the question whether we would actually have enough statistical power to find eQTLs using only 228 samples.

The statistical power (see Section 1.7) in my eQTL study depends on the sample size, the ratio between samples with the minor allele and the major allele, the size of the effect and the desired significance level, which is the probability of a Type I error (false positive). Since the aim of this experiment was only to determine whether my sample size was roughly appropriate, I calculated the power of a comparable t-test to approximate the results I could expect to obtain with the full linear mixed model. I did not consider possible differences between developmental stages for this experiment, but assumed that the effect had the same direction and the same size in all three stages.

Using the R package "pwr" (Champely, 2015) I simulated the statistical power to detect effects of different magnitudes and different MAFs, assuming a sample size of 228 and allowing for a false positive rate of 1 %. Here, the effect sizes are defined as the standardised mean difference, that is, the difference in mean of the two groups divided by their standard deviation. The result is shown in Figure 4.3.

This plot illustrates how my power to detect effects increases with increasing MAF, as the sample sizes between the two genotypes become more balanced and the estimates of mean expression level become more certain. For small effects (0.2) I have little power even at high MAF, suggesting that we will not be able to pick up eQTLs with small effects in this study. However, for larger effect sizes my power looks quite promising. At a MAF of 5 % I already have approximately 50 % power to detect effects of size 0.8, 75 % power to detect effects of size 1, 90 % power to detect effects of size 1.2 and almost 100 % power for effects sizes greater than 1.2. I decided to test only variants above this threshold of 5 % in the eQTL analysis.

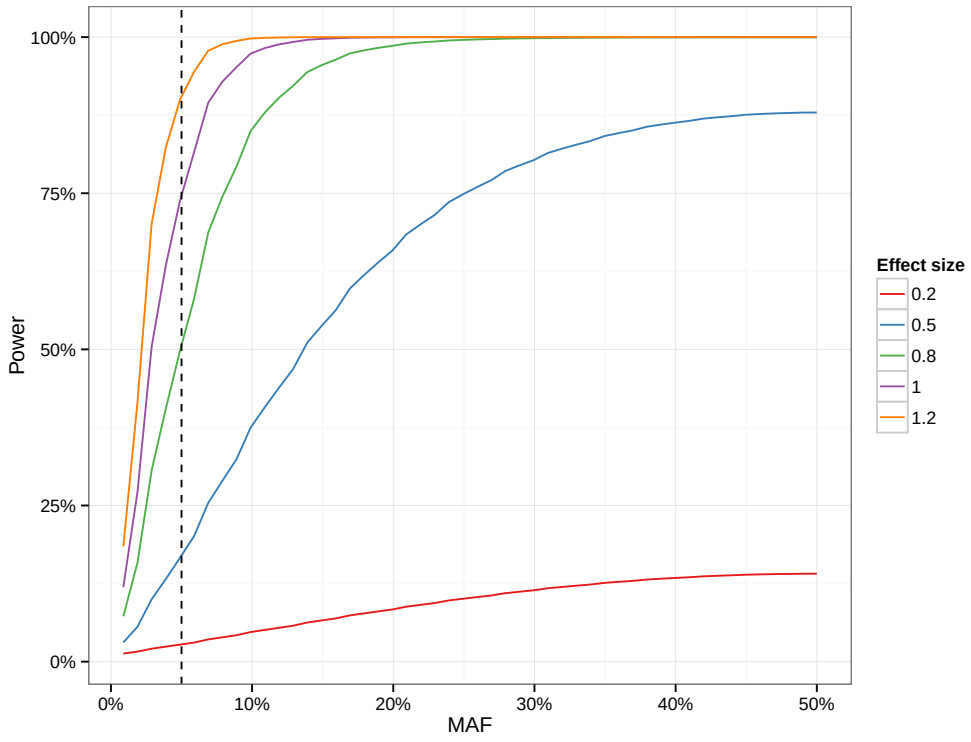


Figure 4.3.: Simulated power to detect effects of different sizes using 228 samples at different MAFs. Dashed line indicates a MAF of 5 %.

4.4. Single-stage eQTL testing

As an initial analysis, I called eQTLs in each of the three developmental stages separately using a univariate linear mixed model similar to the ones described in Bennett et al. (2010) and Lippert et al. (2014):

$$\mathbf{y} = \mu\mathbf{1}_N + \beta\mathbf{x} + \mathbf{g} + \boldsymbol{\psi} \tag{4.2}$$

In which the phenotype is modelled as a sum of the intercept term $\mu\mathbf{1}_N$, a fixed genetic effect term $\beta\mathbf{x}$, a random effect term accounting for the population structure $\mathbf{g}$ and residual noise $\boldsymbol{\psi}$.

I defined this model in LIMIX with the help of Francesco Paolo Casale, a PhD student from the Stegle group and one of the co-authors of the software, using the following variables:

- $\mathbf{y}$: A vector of phenotypes (gene expression levels of the given gene) for all

76 individuals in the given developmental stage (dimension 76).

- μ: The intercept parameter (scalar).
- $\mathbf{1}_N$: A vector of ones (dimension $N = 76$).
- $\mathbf{x}$: Vector of minor allele dosages for each of the 76 lines at the given variant, with 0 being a homozygous major genotype, 1 being a heterozygous minor/major genotype and 2 being a homozygous minor genotype (dimension 76).
- β: The effect size of the genotype at the given SNP in the given developmental stage (scalar).
- $\mathbf{g}$: A random effect term for population structure (dimension 76). Modelled as following a multivariate normal distribution $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K}$ is the genetic relatedness matrix calculated from the full genotype matrix.
- $\boldsymbol{\psi}$: A random effect term representing residual noise (dimension 76). Modelled as normally distributed, independent and identically distributed (i.i.d.) noise: $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix.

I could then test for association between a gene and a variant by comparing the fit of the alternative model with $\beta \neq 0$ and the fit of the no-effect model with $\beta = 0$ using a log-likelihood ratio test in LIMIX. In this way, I tested the set of proximal candidate variants for each gene, obtaining a p-value p and effect size β for each variant.

To account for multiple testing within each gene-proximal region I calculated an empirical p-value e for each uncorrected p-value p (see Section 1.7.1) as follows. I performed 10,000 permutation experiments, randomly permuting the vector of genotypes $\mathbf{x}$ of each variant, assigning a random set of genotypes to each of the samples but keeping the same number of minor, heterozygous and major genotypes. I then re-tested all variants again and determined the minimum uncorrected p-value m_i between all of them.

After performing 10,000 such random permutations for a given gene I obtained a vector of the 10,000 minimal p-values $\mathbf{m} = \langle m_1, \ldots, m_{10000} \rangle$. I then counted how many p-values in this empirical null distribution were lower than the uncorrected p-value p and calculated the empirical p-value e as this count plus one, divided by 10,001: $e = (\text{count}(\mathbf{m} < p) + 1) \cdot 10001^{-1}$.

I applied this procedure to test for proximal eQTLs around each expressed gene in each of the developmental stages separately. In total I tested 10,163 genes at 2–4 h, 10,290 genes at 6–8 h and 10,307 genes at 10–12 h.

4.4.1. Single-stage eQTL results

At an empirical p-value threshold of $\alpha = 1\,\%$ I found at least one eQTL for 2,462 genes at 2–4 h, 2,404 genes at 6–8 h and 2,450 genes at 10–12 h (Table 4.1).

Stage	Genes tested	Genes with eQTLs
2–4 h	10163	2462
6–8 h	10290	2404
10–12 h	10307	2450

Table 4.1.: Number of eQTLs and genes with eQTLs found with the single-stage eQTL tests.

As a first naive approach to estimate how many eQTLs were shared between all three developmental stages, I overlapped these three sets of genes with eQTLs. I ignored the exact location of the identified variant in this overlap to account for some random fluctuations in the data. The resulting Venn diagram is shown in Figure 4.4.

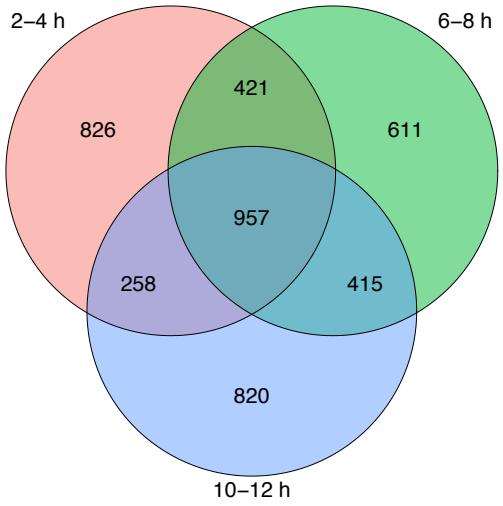


Figure 4.4.: Venn diagram of genes with at least one eQTL in the three different developmental stages.

As expected, there was quite a large number of genes with an eQTL in all three

stages (957), suggesting that the regulatory regions of many genes were being used in all developmental stages that I analysed. In addition, the overlap between pairs of two stages behaved as expected, with the overlap between 2–4 h and 6–8 h and between 6–8 h and 10–12 h being almost twice as large as the overlap between 2–4 h and 10–12 h. This reflects the systematic changes in the developmental transcriptome during embryogenesis, resulting in more similarity between 2–4 h and 6–8 h than between 2–4 h and 10–12 h. However, the amount of genes that only had an eQTL at one of the stages was surprisingly large. Between 25 % and 34 % of genes appeared only to have an eQTL at a single developmental stage.

Upon closer inspection of the data, it became clear that this did not accurately reflect the true situation. In particular, there were many cases where there actually seemed to be an effect in all three developmental stages, but only in one of them did the effect pass the threshold for significance, giving the impression that this eQTL was only active in one stage. This effect was particularly common for genes that were expressed at different levels in the three stages, with the eQTL being much more likely to be detected in the stage with the strongest expression. This is an expected consequence of the single-stage testing approach but makes it challenging to derive meaningful conclusions from comparisons between the three single-stage data sets. I thus moved on to a more rigorous approach that solves this problem by modelling the gene expression levels from all three developmental stages at the same time.

4.5. Multi-stage eQTL testing

Testing for eQTLs with single-stage univariate mixed models had already shown that it was feasible to find eQTLs using our data set. However, it became clear that it was not suitable for differentiating between eQTLs that were active in all three developmental stages and those that were specific to one of them.

In addition to univariate mixed models, LIMIX also allows the specification of multivariate mixed models, which enabled me to model the expression levels at all three developmental stages simultaneously. Using a multivariate model, I could not only test for stage-specific effects in a more accurate way, but could also increase the statistical power by considering the data from all 228 samples together.

When modelling a random effect in a multivariate linear mixed model, one needs to consider the covariance both between the different traits as well as between the

different samples. In my case, I am studying three different developmental stages (traits) and 76 different DGRP lines (samples), which results in a 228×228 covariance matrix that can be factorised into two separate matrices: the trait covariance matrix $\mathbf{C}$ (3×3), which describes the correlation of gene expression levels in each line between the developmental stages induced by the random effect, and the sample covariance matrix $\mathbf{R}$ (76×76), which describes the correlation in gene expression levels in the same developmental stage between lines. The trait covariance matrix and the sample covariance matrix can be combined to yield the full covariance matrix using the Kronecker product: $\mathbf{V} = \mathbf{C} \otimes \mathbf{R}$.

The Kronecker product $\mathbf{A} \otimes \mathbf{B}$ represents a multiplication of the matrix $\mathbf{A}$ with dimensions $m \times n$ with each element of the matrix $\mathbf{B}$ with dimensions $k \times l$ and then stacking the results to form a new matrix with dimensions $mk \times nl$. For example:

$$\begin{bmatrix} -1 \\ 3 \end{bmatrix} \otimes \begin{bmatrix} 10 & 15 \\ 20 & 5 \end{bmatrix} = \begin{bmatrix} -10 & -15 \\ -20 & -5 \\ 30 & 45 \\ 60 & 15 \end{bmatrix}$$

Using this factorisation only the small trait covariance matrix needs to be estimated for each gene and each random effect in the model, while the sample covariance matrix is only estimated once.

To define the multivariate models I used the same variables as described above, with the following additions and changes:

- $\mathbf{y}_{\{1,2,3\}}$: Vectors of gene expression levels in each of the three developmental stages (dimension 76 each). These are stacked to make a vector of dimension 228.
- $\mu_{\{1,2,3\}}$: Separate intercept parameters for each of the developmental stages (scalar).
- β_c: Estimate for the effect size of the genotype at the given SNP, common to all three stages (scalar).
- $\beta_{\{1,2,3\}}$: Estimate for the stage-specific effect size of the genotype at the given SNP in addition to the common effect (scalar).
- $\mathbf{g}$: A random effect term for population structure (dimension 228), the covariance matrix of which factors into a trait and a sample covariance

matrix as described above: $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_g \otimes \mathbf{K})$, where $\mathbf{C}_g$ is the covariance matrix of the developmental stages attributable to genetic factors and $\mathbf{K}$ is the genetic relatedness matrix.

- $\boldsymbol{\psi}$: Residual noise (dimension 228) accounting for covariance between traits but assuming independence between samples: $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_n \otimes \mathbf{I})$, where $\mathbf{C}_n$ is the residual covariance matrix of the developmental stages attributable to non-genetic factors and $\mathbf{I}$ is the identity matrix.

Using these symbols and conventions I defined the following multivariate models, similar to the ones described in Korte et al. (2012) and Lippert et al. (2014):

- The common effect model:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \otimes \mathbf{1}_N + \begin{bmatrix} \beta_c \\ \beta_c \\ \beta_c \end{bmatrix} \otimes \mathbf{x} + \mathbf{g} + \boldsymbol{\psi} \tag{4.3}$$

  This models the case that the expression level of the gene is explained by a genetic effect that is the same in all three developmental stages, plus the intercept term and random effects as described above.

- The specific effect models:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \otimes \mathbf{1}_N + \begin{bmatrix} \beta_c + \beta_1 \\ \beta_c \\ \beta_c \end{bmatrix} \otimes \mathbf{x} + \mathbf{g} + \boldsymbol{\psi} \tag{4.4}$$

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \otimes \mathbf{1}_N + \begin{bmatrix} \beta_c \\ \beta_c + \beta_2 \\ \beta_c \end{bmatrix} \otimes \mathbf{x} + \mathbf{g} + \boldsymbol{\psi} \tag{4.5}$$

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \otimes \mathbf{1}_N + \begin{bmatrix} \beta_c \\ \beta_c \\ \beta_c + \beta_3 \end{bmatrix} \otimes \mathbf{x} + \mathbf{g} + \boldsymbol{\psi} \tag{4.6}$$

  These models can be used to test whether there is a genetic effect specific to one of the stages in addition to the common effect.

Similar to the single-stage model, I could now test the common-effect model (where $\beta_c \neq 0$) against the no-effect model (where $\beta_c = 0$) to look for a variant that had the same effect at all three developmental stages. However, in addition to this I could then carry out three more tests, calculating the p-value for the log-likelihood ratio test of each of the three specific models against the common-effect model. This allowed me to determine whether there was an additional effect in one of the stages that went beyond the common effect.

I now applied these tests to our data set as before, testing all biallelic variants with MAF $> 5\%$ in a 50 kb region around each gene and calculating an empirical p-value to correct for multiple testing per gene. I only tested the 10,094 genes which were expressed in all three developmental stages, using the criteria described in Section 3.5. An overview over the parameters of the multi-stage eQTL test is shown in Table 4.2.

Genes	Tested Variants	Tests per gene
10094	1757540	1579.92

Table 4.2.: Multi-stage eQTL testing parameters: Number of unique genes tested, number of unique tested variants and mean number of tests performed per gene.

The number of tests performed per gene differed between genes, due to the varying density of annotated variants across the genome. A histogram of the number of tested variants per gene is shown in Figure 4.5. The minimum number of variants I tested for a gene was 103, the maximum was 5,601.

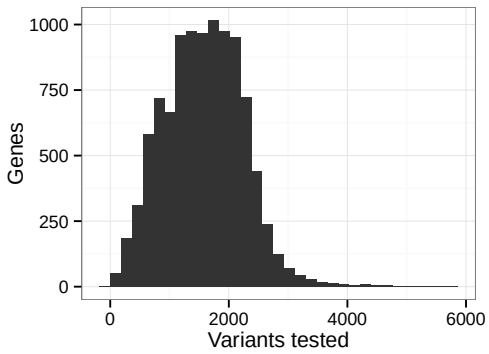


Figure 4.5.: Histogram of the number of variants tested with each gene in the multi-stage eQTL testing.

4.6. Processing of multi-stage eQTL testing results into eQTL sets

The output of the eQTL calling step can be conceptualised as a set of $V \times T$ matrices for each gene, where V is the number of variants and T is the number of tests that were performed (1 common test + 3 specific tests = 4 tests). These matrices contain all the important output for each log-likelihood test with each variant, including the uncorrected p-value p, the empirical p-value e and the effect size β. The challenge now was to process this data in a way that allowed me to derive biologically meaningful and interpretable results from it.

As a first step, I generated four sets of eQTLs that passed our empirical p-value threshold of $\alpha = 1\,\%$, one for each of the four tests I performed[1]. The total number of eQTLs as well as the amount of unique genes with at least one eQTL for each of these four sets is shown in Table 4.3.

Test	eQTLs	Genes with eQTLs
Common effect	49649	3139
Specific effect at 2–4 h	3960	447
Specific effect at 6–8 h	906	186
Specific effect at 10–12 h	2229	335

Table 4.3.: Number of gene eQTLs and unique genes with eQTLs found with the multi-stage eQTL tests.

From these full eQTL sets, I generated four summarised eQTL sets which contained the eQTL with the lowest uncorrected p-value (the lead eQTL) for each gene. I called these the "overview" eQTL sets.

The vast majority of the eQTLs that I found had a common effect. However, it is important to highlight two aspects of the eQTL testing which mean that an eQTL will be more likely to have a common effect than a specific effect:

First, if the effect sizes associated with a variant differ between the three developmental stages, the common effect size β_c will essentially be an average of the three true effect sizes. Thus, if the effect is strong in one of the stages and weak but not opposite in the other stages, the common effect test may still find an association, but with a decreased effect size. Second, for a specific effect to be identified, the model with the specific effect has to fit the data significantly

[1] This threshold corresponds to a false-discovery rate (FDR) of approximately 10 % across all tested genes and models.

better than the common effect model, which is a stringent criterion.

Together, these two aspects of the test design may result in cases where a specific effect is "split up" into a weak common effect, which underestimates the effect size, and a weak specific effect, which corrects this underestimation. Either or both of these effects may fail to pass the significance threshold, which can result in me either classifying a specific effect as a common effect or missing the association completely.

Interestingly, I also found much fewer eQTLs specific to 6–8 h than to the other two time points. A possible explanation for this may be that there are some regulatory mechanisms that are most active around the first time point in this study and then become less used over time, while the usage of other mechanisms is low at first but then increases over time. This would be in line with the fact that 2–4 h covers both the syncytium stage as well as the maternal-zygotic transition (MZT), which are associated with direct regulation through protein gradients and miRNAs, while at 10–12 h gene regulation occurs more often through indirect signalling pathways.

Consequently, 2–4 h and 10–12 h could be expected to be quite different from each other, while 6–8 h as the midpoint would share some features with 2–4 h and some features with 10–12 h and would thus have fewer unique eQTLs. However, it is also possible that this difference is caused, at least in part, by the experimental design, in which most of the samples from 6–8 h were sequenced before the other samples. This may have increased the technical variance of the 6–8 h samples, which would have resulted in a lower statistical power to detect effects at 6–8 h and thus fewer stage-specific eQTLs. I will comment on this aspect of the experimental design in Chapter 8.

4.6.1. eQTL clouds

As described in Section 1.9, there is a very low degree of linkage disequilibrium (LD) in the DGRP. In theory, this low LD, coupled with the fact that the DGRP lines have been fully sequenced, may allow for the identification of the exact causal locus for many genes. However, while the genome-wide LD is low, there are still some regions in the genome where variants are in high LD with each other. If any one of these variants has an effect on a gene's expression level, all the other variants in LD will be associated with the expression level as well. This effect can lead to large numbers of variants being associated with a single gene. Figure 4.6 shows that, while the number of associated variants per gene was generally low,

there are some genes that had more than a hundred different variants associated with their expression level.

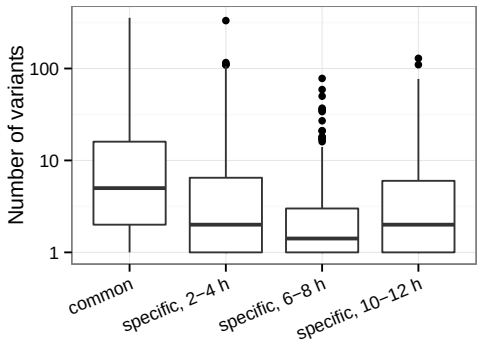


Figure 4.6.: Boxplot of the number of associated variants per gene found with each of the four tests. Genes without any significant associations not shown.

In order to determine whether these associations were caused by LD or represent separate genetic effects I could have performed a conditional analysis: For each gene with more than one significant association, I could compare a model with only the lead variant as a fixed effect to a model with the lead variant plus a second variant as fixed effects. If the secondary variants were merely in LD with the lead variant, the log-likelihood ratio test of the two-variant model against the single-variant model should not result in a significant p-value. If, however, there were actually multiple effects acting independently, the model with two variants should fit the data better than the single-variant model, resulting in a significant p-value. This procedure could then be continued with a possible third variant, and so on, until there are no more significant associations. However, due to the relatively low sample size in our study, I did not attempt to perform this analysis, since I would have been unlikely to have enough statistical power to test for the second variant in almost all of the cases.

Even in cases of high LD the causal variant should be the one most strongly associated with the expression levels and thus be listed as the lead variant. However, due to the stochasticity involved in gene expression level measurement and eQTL testing, sometimes another variant may randomly have been more strongly associated, resulting in that variant being listed as the lead variant in the "overview" set.

In order to account for this uncertainty, I generated four more sets of eQTLs that did not just contain the lead variant for each gene, but also all other associated

variants with an uncorrected p-value that was within one order of magnitude, similar to the approach used in Ding et al. (2014). This resulted in a set of "eQTL clouds" containing one or more likely causal variants for each gene (Table 4.4).

Test	Genes with eQTLs	Variants per cloud
Common effect	3139	4.17
Specific effect at 2–4 h	447	4.35
Specific effect at 6–8 h	186	3.32
Specific effect at 10–12 h	335	3.94

Table 4.4.: Number of variants in gene eQTL clouds for each of the multi-stage tests.

A boxplot of the number of variants in the eQTL clouds of the different genes is shown in Figure 4.7. While there were still some cases of close to 100 variants associated with a single gene, the numbers decreased across almost all of the cases.

The size of the region encompassed by the variants in the clouds (measured from the leftmost to the rightmost variant) is shown in Figure 4.8. Again, most eQTL clouds only covered a small region, indicating that they may only contain the causal variant and a few variants close-by that were in LD with it. However, there are cases where the region covered by the variants is much larger. The fact that eQTLs are so far away from each other could either be caused by very large regions of LD, or indicate the presence of multiple uncorrelated eQTLs acting independently on gene expression.

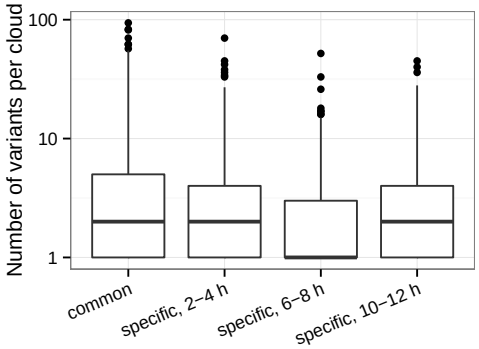


Figure 4.7.: Boxplot of the number of variants per gene within the eQTL cloud found with each of the four tests. Genes without any associated variants not shown.

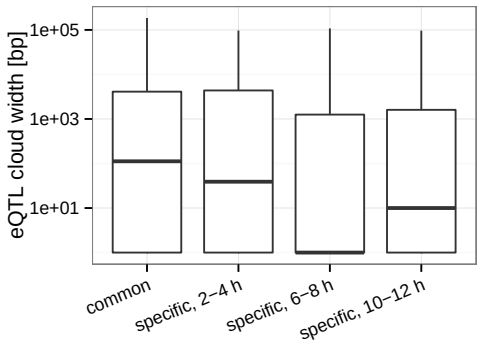


Figure 4.8.: Boxplot of the width of eQTL clouds found with each of the four tests in bp. Genes without any associated variants not shown.

4.6.2. Interpreting stage-specific effects

One of the big advantages of the multi-stage eQTL testing over the single-stage eQTL testing was that it would allow me to determine which eQTLs were actually active in all three developmental stages and which ones were stage-specific in a statistically rigorous way. However, due to the way I tested for common and specific effects, I could not easily deduce this information from the eQTL calls. In particular, I noticed that there were quite a few cases where there was a common effect in one direction, which was then counteracted by a specific effect in the opposite direction. This meant that, while the eQTL had been called as specific in a given stage X, the eQTL was actually active in the other two stages Y and Z and specifically *not* active in X.

In order to handle these cases appropriately and extract a set of eQTLs that were truly active at only a single stage, I devised a classification procedure, which I applied to all variants that showed a significant association in at least one of four multi-stage tests. This classification procedure made use of the results from the single-stage eQTL testing, and, in particular, the effect sizes I had calculated for each variant/gene pair. I extracted the corresponding effect sizes from the three single-stage tests for each multi-stage eQTL, which gave me an indication of how much of an effect each of the multi-stage eQTLs had in each of the three stages.

For every eQTL, I first checked whether it had a significant common effect. If it did, I immediately classified it as a common eQTL, regardless of the other tests. This merged the eQTLs active in two stages with the eQTLs active in three stages, but greatly simplified the classification. Most eQTLs were common, indicating

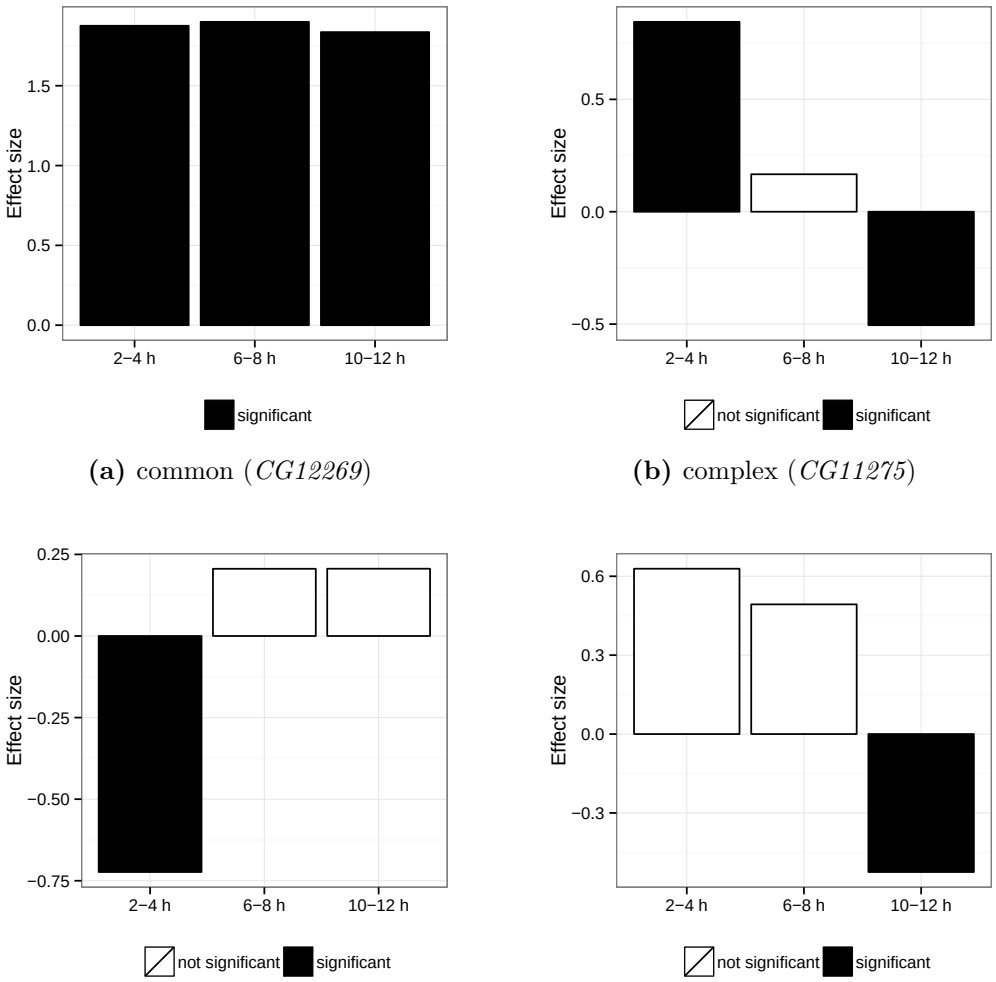
that most effects were indeed shared across at least two stages. An example of the single-stage effect sizes of a common eQTL in the three stages is shown in Figure 4.9a.

If an eQTL did not have any common effect, but had more than one specific effect, I classified the eQTL as "complex". These are cases where the eQTL may not just be active in one stage and inactive in the other, but may actually have an opposite effect in two different stages. An example of such a case is shown in Figure 4.9b.

Finally, for all eQTLs that only had a single significant specific effect, I compared the effect sizes that had been estimated for this eQTL in the single-stage eQTL test for the three stages. If the effect size was largest for the stage in which the specific effect was called, I classified it as a single-stage eQTL. This is the group of eQTLs for which I could be confident that they had an effect in only one stage. Figure 4.9c shows the effect sizes of an eQTL that was specific to 2–4 h.

If the effect sizes were not consistent with this being a single-stage eQTL (e.g. the eQTL was called specific to stage X, but its effect size was actually larger in stage Y) I classified this as a weak complex eQTL. An example of this pattern is shown in Figure 4.9d. Manual inspection of some of these weak eQTLs revealed that they were most likely eQTLs that were acting at two of the stages, but fell just below the significance threshold to be called a common effect.

A flow chart of this procedure is shown in Figure 4.10.



(a) common (*CG12269*)

(b) complex (*CG11275*)

(c) single-stage: 2–4 h (*CG1792*)

(d) weak (*CG9577*)

Figure 4.9.: Examples of effect sizes obtained from the single-stage eQTL testing for multi-stage eQTLs from different specificity classes. Gene names shown in parenthesis.

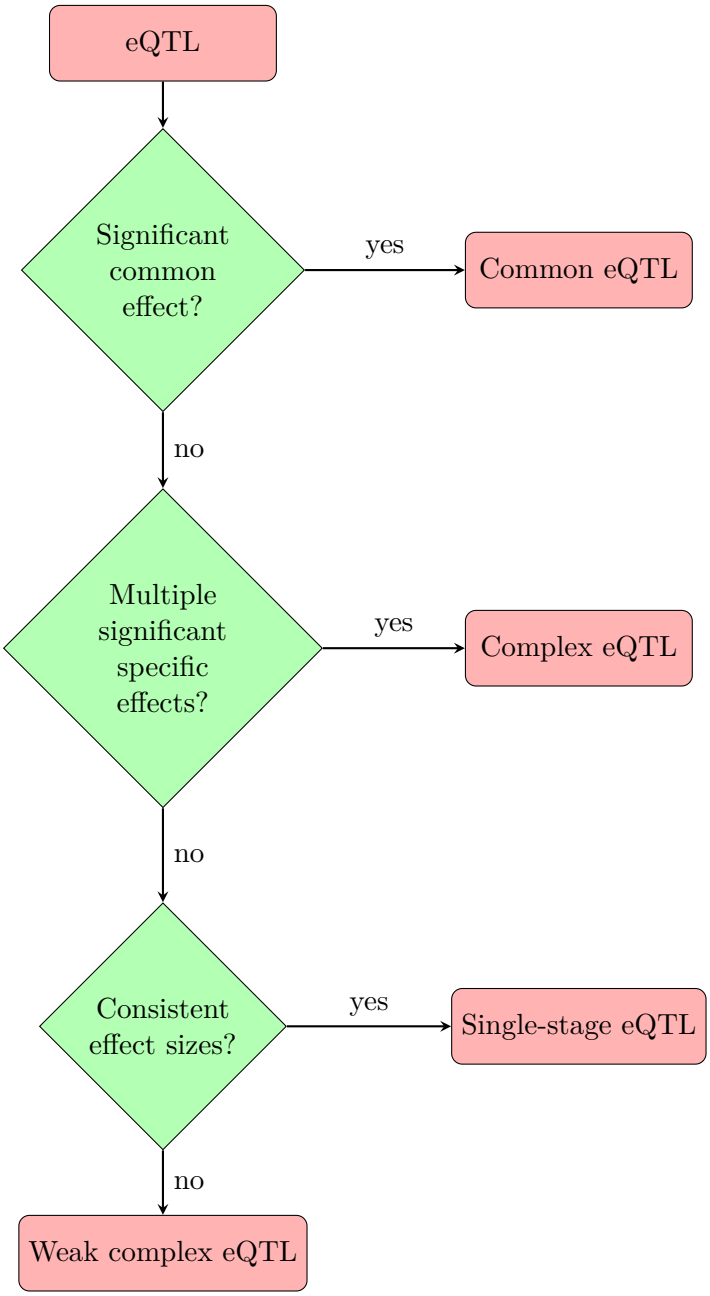


Figure 4.10.: Flow chart illustrating the classification of an eQTL into common, complex, weak complex or single-stage eQTLs.

4.7. Comparison between single-stage and multi-stage eQTL tests

I now compared the results from the single-stage eQTL testing to the results from the multi-stage eQTL testing. Since the multi-stage testing was in theory better powered to detect effects common to all three stages, I first overlapped the set of genes with common multi-stage eQTLs with the set of genes with eQTLs in all three single-stage tests (Figure 4.11).

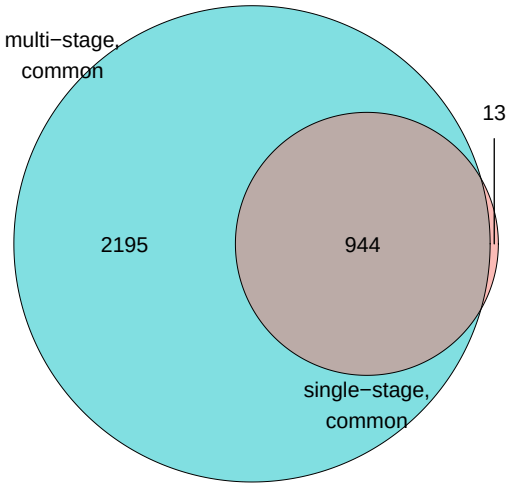
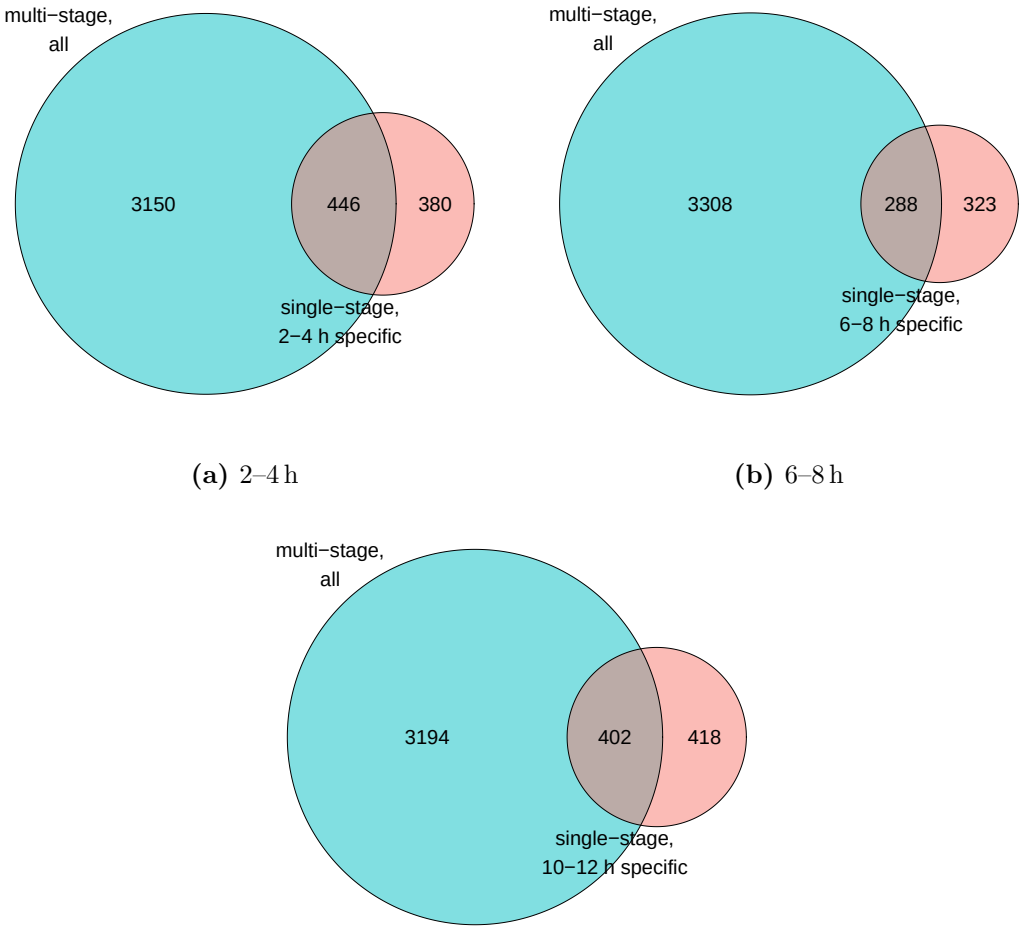


Figure 4.11.: Venn diagram of genes with at least one eQTL common to the three single-stage tests (red) and common in the multi-stage eQTL testing (blue).

Reassuringly, almost all of the common single-stage eQTLs (944/957) were also found with the multi-stage test, confirming that the common-effect test worked as expected. In addition, the multi-stage testing found a large number of additional common eQTLs (2,195), which I had not been able to detect or classify correctly with the single-stage tests. Finally, 13 common eQTLs were found with the single-stage testing but not with the multi-stage testing. This discrepancy may have been caused by random variation in the model fitting and calculation of empirical p-values. In fact, for seven of these eQTLs I found a common association in the multi-stage testing that was close to passing the significance thresholds as well, with an empirical p-value below 0.05. In addition, complex specific effects, with an eQTL having a positive effect in one stage and a negative effect in the other, may have been misclassified as common in the single-stage testing but not in the multi-stage testing.

To compare how many of the single-stage effects were also found with the multi-stage testing, I overlapped the three sets of single-stage eQTLs that were specific to a single time point with the set of all multi-stage eQTLs (Figure 4.12).



(a) 2–4 h **(b)** 6–8 h

(c) 10–12 h

Figure 4.12.: Venn diagram of genes with at least one eQTL specific to the given developmental stage from the single-stage eQTL testing (red) and all eQTLs from the multi-stage eQTL testing (blue).

At 2–4 h, 446/826 (54 %) single-stage eQTLs were also found in the multi-stage testing. Most (345) of these were found to have a common effect in the multi-stage set, indicating that they had previously been misclassified as specific. The remaining 101 eQTLs were in one of the stage-specific eQTL sets. Surprisingly, the remaining 380 single-stage eQTLs from 2–4 h were not found in the multi-stage

eQTL testing at all. Closer inspection of the data for these revealed that almost all of them had been classified as a small common effect plus a small specific effect in the multi-stage eQTL testing, however with neither of them reaching the significance threshold. I observed a similar overlap at the two other developmental stages as well, with many formerly stage-specific effects now being classified as common, but some no longer being found.

This effect, which I described in Section 4.6, is an unfortunate consequence of the test design that appears to have decreased the number of stage-specific eQTLs that I was able to call. However, the multi-stage testing allowed me to identify many more common eQTLs than the single-stage tests. A full four-way overlap of the genes with eQTLs in each of the three single-stage tests as well as in the multi-stage test is shown in Figure 4.13.

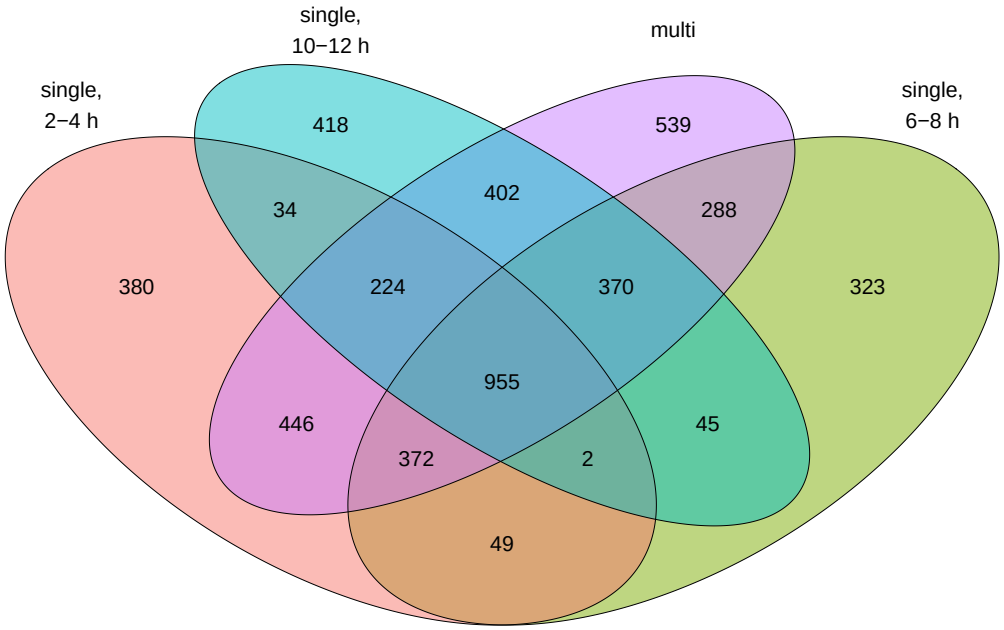


Figure 4.13.: Venn diagram of genes with at least one eQTL in the single-stage eQTL testing (red, blue, green) and at least one eQTL in the multi-stage eQTL testing (purple).

4.8. Comparison to variance decomposition

In Section 4.2 I decomposed the variance of the expression levels of 6,529 genes into different components, assigning them to either additive genetic effects (proximal and distal), interaction effects between genetics and the developmental stage (proximal and distal), the developmental stage itself and noise. My expectation was that I would be more likely to find eQTLs for genes for which the proximal genetic and genetic/development components accounted for a large percentage of the total variance and less likely to find eQTLs when these components were small. In addition, I expected that genes with a large pure genetic component were likely to be common eQTLs, while genes with a large genetic/development interaction component were more likely to be stage-specific eQTLs.

I could now test these assumptions by determining whether I had found at least one proximal eQTL for each of the 6,529 genes. Figure 4.14 shows the percentage of variance explained by the sum of the two proximal genetic components, grouped by whether the gene had an eQTL. As expected, the genes with eQTLs had much larger fractions of their variance explained by genetic effects (Wilcoxon rank sum test, $p = 2.83 \cdot 10^{-232}$).

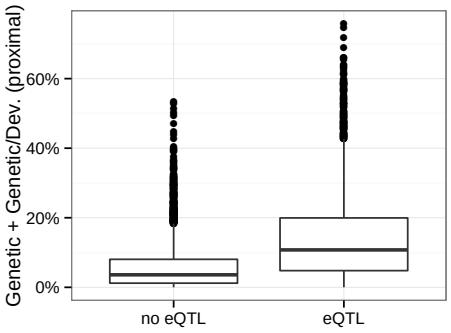
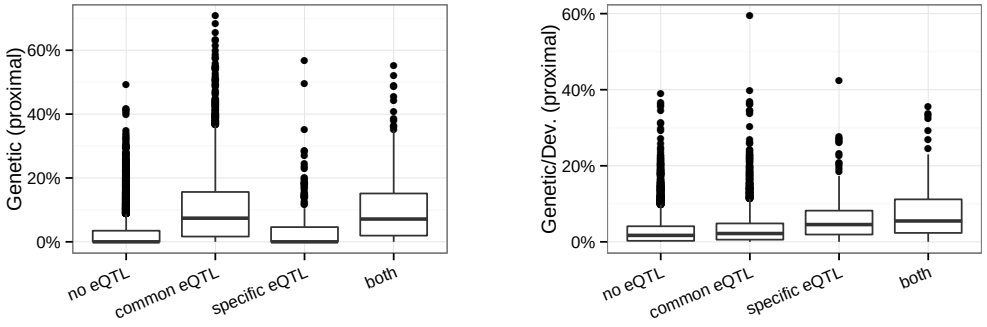


Figure 4.14.: Boxplots of the percentage of variance in gene expression levels explained by proximal genetic components, grouped by whether I found a proximal eQTL for the gene.

To compare the difference between pure genetic effects and genetic/development interaction effects, I further split up the genes by whether they had no eQTL, an eQTL with a common effect, an eQTL with a specific effect or both (Figure 4.15). Genes with common eQTLs had a significantly larger pure genetic component than genes without any eQTL (Wilcoxon rank sum test, $p = 1.26 \cdot 10^{-271}$) and genes with specific eQTLs had a significantly larger genetics/development interaction

component than genes with a common eQTL (Wilcoxon rank sum test, $p = 1.23 \cdot 10^{-42}$).



(a) Genetic **(b)** Genetic/Development interaction

Figure 4.15.: Boxplots of the percentage of variance in gene expression levels explained by pure proximal genetic effects (a) and interaction between proximal genetic and developmental effects (b), grouped by eQTL specificity.

This analysis showed that the results from the variance decomposition agreed with the results from the eQTL testing, despite the different normalisation methods used in the two data sets. However, the agreement was not perfect and there were some genes with large genetic components for which I did not find an eQTL. For example, I was unable to find any eQTL for 40 (12 %) of the 323 genes for which I had estimated that more than 25 % of the variance could be attributed to the proximal genetic component.

It is likely that this difference is because I estimated the variance explained by all proximal variants as a group in the variance decomposition, while in the eQTL testing I tested each variant individually for association. This means that if the expression level of a gene was influenced by several weak proximal genetic effects, I may have been able to detect the sum of these effects in the variance decomposition, but may not have had enough statistical power to detect an effect of any single variant.

4.9. 3′ isoform eQTLs

As described in Chapter 2, I initially calculated the expression levels of individual 3′ end peaks, which I then summed up to form the gene expression levels. While this gene expression level is the trait in which changes are more likely to have an effect on the organism, I was also curious about what eQTLs for individual

peaks could tell me about the mechanism and regulation of alternative 3′ isoform usage. Thus, in addition to calling eQTLs on the total gene expression level, I also looked for eQTLs that affected only a single 3′ end peak.

For this, I applied the multi-variate eQTL testing procedure described in Sections 4.5 and 4.6 to the set of 3′ Tag-Seq peak expression levels. I tested a total of 31,913 peaks for association with variants in a 50 kb region upstream and downstream of the peak, for an average of approximately 1,392 variants per peak (Table 4.5).

Peaks	Tested Variants	Tests per peak
31913	1762906	1392.12

Table 4.5.: Multi-stage eQTL testing parameters: Number of unique peaks tested, number of unique tested variants and mean number of tests performed per peak.

As before, I called four different sets of eQTLs, testing for both common effects and effects specific to one of the time points. The resulting total number of 3′ isoform eQTLs is shown in Table 4.6.

Test	Peaks with eQTLs	Variants per cloud
Common effect	10387	4.02
Specific effect at 2–4 h	822	3.63
Specific effect at 6–8 h	380	2.84
Specific effect at 10–12 h	669	3.32

Table 4.6.: Number of variants in peak eQTL clouds for each of the multi-stage tests.

I will return to this set of 3′ isoform eQTLs in Chapter 6 to study cases of 3′ isoform-specific eQTLs and alternative polyadenylation.

5. Quality-control of eQTLs

In this chapter I begin to explore the sets of eQTLs I found in Chapter 4, particularly with regard to technical confounders. In order to identify false positives that may have been caused by the 3′ Tag-Seq protocol, I estimate the concordance between eQTLs and gene expression levels measured with standard RNA-seq. Based on the differences I observe, I add additional filtering steps which aim to remove many of these false positives. Finally, I show examples of eQTLs, including two associated with the classic *Drosophila* genes *white* and *Dichaete*, and describe how I validated the set of eQTLs using *in situ* hybridisation and previously published data.

5.1. Validation of 3′ Tag-Seq eQTLs with RNA-seq

In Section 2.8, I described how I compared the gene expression levels obtained from 3′ Tag-Seq with the ones obtained from RNA-seq. I observed that, in general, the mean gene expression levels were correlated well across samples, with Spearman's $\rho = 0.90$. However, I also observed that for some genes the expression levels differed between RNA-seq and 3′ Tag-Seq, suggesting that one of the measurements was incorrect. Thus, I wanted to ensure that the eQTLs found with 3′ Tag-Seq were not caused by protocol-specific biases, and would replicate when gene expression levels were measured with other methods.

Since the promoter region at the 5′ end of genes is essential for the regulation of transcription (see Section 1.3.1), an enrichment of eQTLs around the 5′ end of their gene would be expected. To check for such an enrichment, I calculated the location of each lead gene eQTL with a common effect relative to its associated gene. Locations within each gene body were scaled by the gene length to allow for comparison between them. The resulting plot is shown in Figure 5.1.

As expected, there was an enrichment of eQTLs at the 5′ end of genes. However, the largest aggregation of eQTLs was observed at the 3′ end. While such an enrichment of eQTLs around the 3′ end of genes had also been shown before in

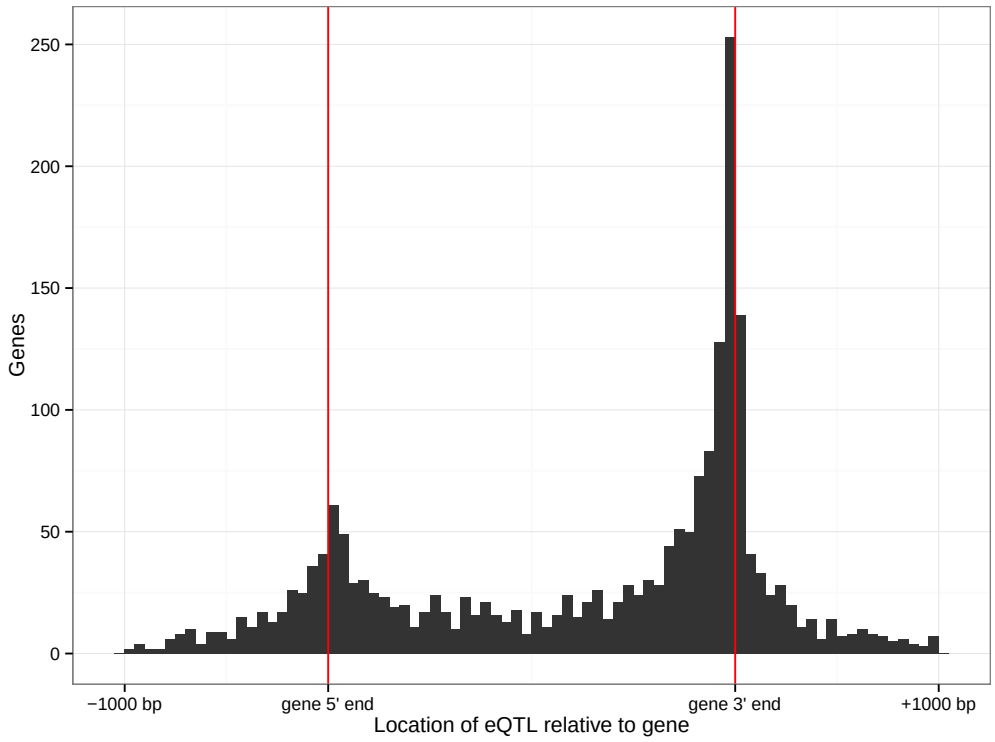


Figure 5.1.: Position of lead gene eQTLs (common effect) relative to their associated gene. Locations inside the gene body in fractions of gene length, locations outside of the gene body in raw bp. Red lines denote location of the 5′ and 3′ end of the gene.

the DGRP (Massouras et al., 2012) as well as other eQTL studies (Tung et al., 2015), the strength of the signal in my data was larger than expected.

In fact, a similar enrichment of eQTLs around the 3′ end of genes had previously been found in humans (Veyrieras et al., 2008). However, in a follow-up study using RNA-seq instead of microarrays, the authors found that this enrichment was no longer present (Veyrieras et al., 2012). This lead them to conclude that the 3′ enrichment was almost entirely an artefact caused by their use of microarrays with probes at the 3′ end of transcripts, which meant that any eQTL affecting the usage of the last exon of a gene would look like an eQTL for the whole gene.

In theory, 3′ Tag-Seq should not exhibit this problem since I called 3′ end peaks *de novo* without any prior expectation of where the 3′ end of a transcript would be (see Chapter 2). When calculating the total gene expression level, I considered peaks at annotated transcript ends as well as inside the gene body, which meant that a mere shift in the location of a 3′ end should not have resulted in a change

in the estimated gene expression level. However, if a 3′ end was located more than 2 kb away from the annotated end I may have assigned it to another gene, which would have led to false positives.

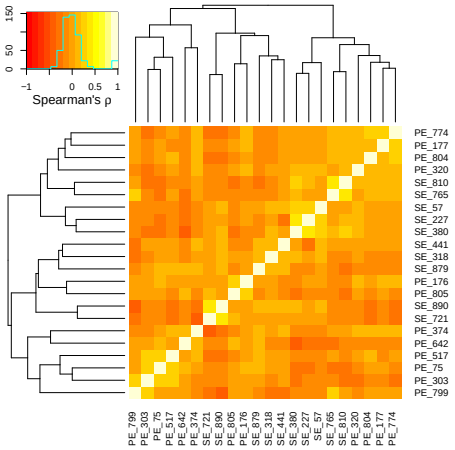
In addition, false positive eQTLs caused by variants associated with differential mappability (as described in Section 2.2) are likely to be located inside or very close to the 3′ Tag-Seq peaks, resulting in an enrichment at the 3′ end of the gene. I thus needed to convince myself of the validity of the eQTLs, and in particular show that the eQTLs could be reproduced using RNA-seq as an alternative method of estimating gene expression levels.

To this end, I checked the effect sizes predicted by my eQTLs for concordance with the effects I could observe in our RNA-seq data. As described in Section 2.8, this data set consisted of 22 libraries of standard Illumina stranded poly(A)$^+$ RNA-seq, which were obtained from the same samples of RNA that had been used for the 3′ Tag-Seq measurements at 10–12 h. It is important to stress that these samples, while well suited to account for differences in sequencing protocols, did not represent independent biological replicates. Thus, this analysis did not allow me to identify false positives caused by other sources of errors, such as issues with the quality of the genotype data, population structure, or problems during sample preparation or collection.

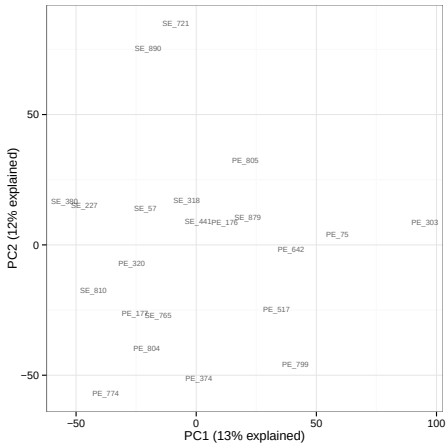
To remove any unwanted population structure and batch effects due to the different sequencing runs, I applied similar normalisation steps to the RNA-seq expression levels as I had applied to the 3′ Tag-Seq data (centring and scaling, quantile-normalisation, PEER with 5 hidden factors).

A heatmap of the final data set and a plot of the samples projected onto the first two principal components from a PCA are shown in Figure 5.2. As for the 3′ Tag-Seq data, the normalisation had successfully removed large-scale experimental batch and population structure effects, resulting in a data set with little structure, well suited for eQTL testing. The two different sequencing methods used (paired-end/PE and single-end/SE) did not form any strong clusters, confirming that they could be combined without further adjustments.

Using this normalised set of RNA-seq expression levels I now attempted to replicate the effects predicted by my 3′ Tag-Seq eQTLs, using the set of lead gene eQTLs with a common effect as my test set. For this, I calculated the difference in RNA-seq expression levels between samples with the homozygous major and the homozygous minor genotype for each eQTL. Based on this difference, I could then classify each eQTL by whether the direction of the change in RNA-seq medians



(a) Clustered correlation between samples



(b) Samples projected onto their first two principal components

Figure 5.2.: Diagnostic plots of RNA-seq samples from the 10–12 h time point.

was concordant with the direction predicted by the 3′ Tag-Seq effect size. I only applied this test to eQTLs where I had obtained RNA-seq expression levels for at least two lines homozygous for each the major and the minor allele, to avoid incorrect calls caused by single outliers.

I observed a concordance rate of 76 % for the eQTLs overall, which was significantly better than the rate of 50 % I would have expected at random (exact binomial test, $p = 2.36 \cdot 10^{-156}$). This indicated that the eQTLs were generally predictive of the gene expression level, even when measured with a different method. The fact that not all of the eQTLs were concordant could at least partially be explained by experimental variation and the fact that the number of samples in the RNA-seq data was lower, leading to higher noise.

However, I also observed that the concordance of eQTLs differed based on their location relative to the gene, with a concordance rate of 86 % for eQTLs at the 5′ end of genes but only 69 % for eQTLs at the 3′ end of genes (Figure 5.3).

Furthermore, Figure 5.4 shows the relative location of a high-confidence set of 542 eQTLs where the eQTL could be validated in the RNA-seq data. I defined this set as all eQTLs where the RNA-seq expression level was significantly different using a Wilcoxon test between the major and minor allele (unadjusted $p < 0.05$) and the sign of the difference in medians was concordant with the eQTL effect direction. Comparing this to the locations of all eQTLs (Figure 5.1) reinforces the notion that the eQTLs around the 5′ end were much more likely to be valid

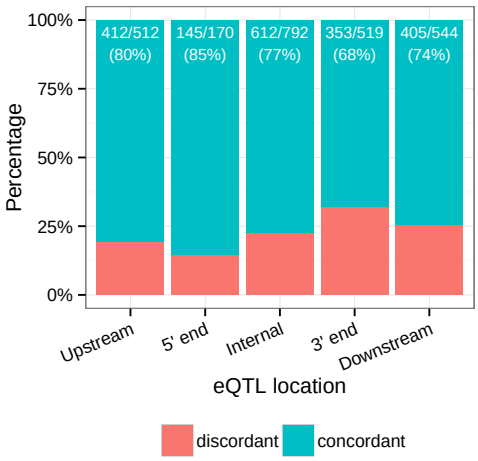


Figure 5.3.: Fraction of concordant and discordant eQTLs grouped by their location relative to their associated gene.

than the ones around the 3′ end. While the peak at the 3′ end was much larger before, the height is now almost exactly equal between the 3′ and 5′ end.

This difference in quality between the 3′ and the 5′ end suggested that there was still an effect resulting in more false positives around the 3′ end of the gene, despite my efforts to account for the mappability of peak regions. Thus, I explored criteria by which I could filter my eQTLs to improve their quality, using the concordance as an estimate of their true positive rate.

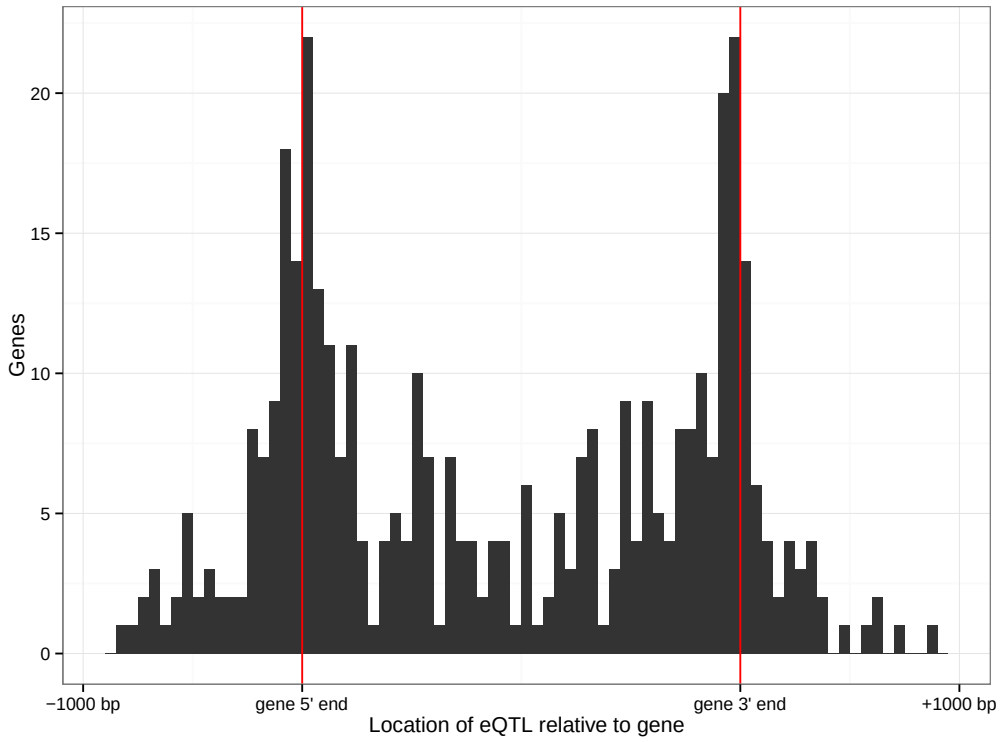


Figure 5.4.: Position of high-confidence lead gene eQTLs (common effect) relative to their associated gene. Locations inside the gene body in fractions of gene length, locations outside of the gene body in raw bp. Red lines denote location of the 5′ and 3′ end of the gene.

5.2. Filtering of eQTL sets

5.2.1. Estimating the mappability of the genome

Since mappability was still the most likely culprit for possible false positive eQTLs, I looked for a way to identify regions of low mappability in the genome. Jacob Degner, a post-doctoral fellow from the Furlong group at EMBL-Heidelberg, had encountered a similar question in a project of his and kindly shared his data with me. For each nucleotide of the *D. melanogaster* genome, Jacob had estimated whether this nucleotide was mappable in each individual DGRP line, based on a mapping of the genome DNA-seq data provided by the DGRP back to the reference genome. Using this data set, I calculated aggregate mappability statistics for each 3′ Tag-Seq peak, including the minimum and mean mappability in the peak region.

In addition, for each DGRP line, I calculated the fraction of mappable bases in all 3′ Tag-Seq peaks associated with each gene separately. For each eQTL, I could then test for a significant difference in mappability between the major and the minor alleles using a Wilcoxon rank sum test. I also tested whether the mappability of a line was associated with the estimated 3′ Tag-Seq gene expression level of that line at each developmental stage, by calculating Spearman's ρ between the mappability and the expression level estimates.

5.2.2. Selection of filtering parameters

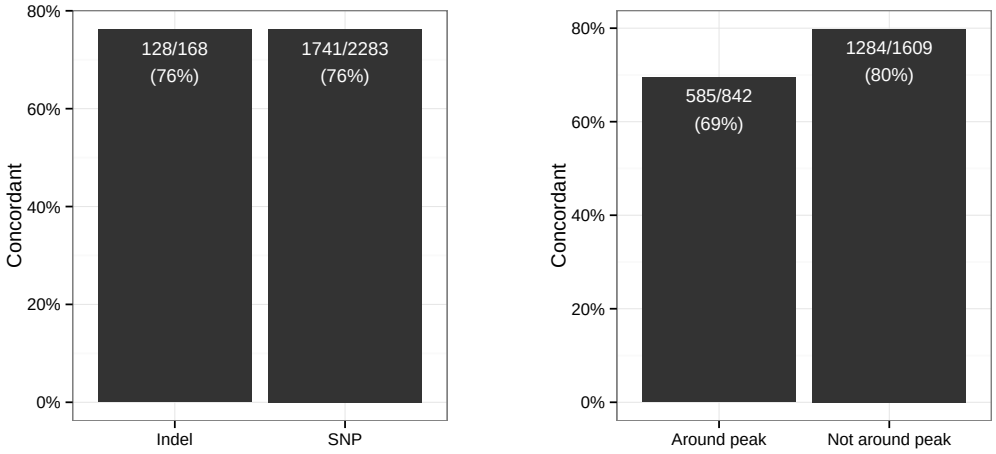
Using the RNA-seq concordance data described in Section 5.1 as a quality measure I now tested the effect of the mappability parameters as well as additional filtering criteria on the quality of the eQTLs. My aim was to find a set of eQTLs that had a good concordance rate, while retaining as many eQTLs as possible. I used the set of lead eQTLs from the common effect test on gene expression levels to optimise these parameters, but then extended them to all other eQTL sets as well.

The parameters I tested for their effect on concordance were:

- Properties affecting mappability of a gene's 3′ Tag-Seq peaks:
  - How large was the fraction of heterozygous variants in the peaks? Heterozygous variants, which were still present in the lines despite their inbreeding, were not accounted for during genome personalisation.
  - How many variants were there in these peaks in total? Highly variable regions were more likely to contain unidentified and rare variants, which were not accounted for during genome personalisation.
  - How long were the variants in these peaks on average? Long indels were more likely to cause mappability problems than short indels or SNPs.
  - What was the estimated minimum level of mappability in the peaks?
  - What was the estimated mean level of mappability in the peaks?
  - What was the p-value of the association test between mappability and genotype in the peaks? The genotype at the eQTL should have been independent of the mappability of the peak.
  - What was the minimum p-value of the association test between mappability and expression level in the peaks? The expression level should have been independent of the mappability of the peak.

- The presence of certain genomic sequences close to the variant, which may have been indicative of problems during the 3′ Tag-Seq protocol that may have led to artefacts.
  - A run of 8, 10, or 12 A nucleotides, which may have been associated with oligo-dT mispriming.
  - Parts of the adaptors and PCR primers used in the 3′ Tag-Seq protocol.
- Properties of the lead variant:
  - Was the variant an indel? Indels may have had larger effects on genes and their regulatory regions, but may also have caused bigger problems in mappability.
  - Was the variant located in, or close to, a 3′ Tag-Seq peak? Variants inside the 3′ Tag-Seq peak were more likely to be associated with mappability than variants further away.
  - What was the MAF of the variant? A larger MAF meant I had higher statistical power.
- Properties of the gene:
  - Did the gene only have a single 3′ Tag-Seq peak or more? A complex 3′ isoform structure may have introduced artefacts into the gene expression level estimates.
  - How long was the gene?
  - How highly expressed was the gene? The variance of expression levels caused by technical factors decreases the more highly expressed a gene is, increasing power.
- Properties of the eQTL:
  - How many other associated variants did the gene have in the eQTL cloud? A large number here may have been related to residual population structure or long-range LD.
  - How strong was the effect size and the $\log_2$(fold-change)? A stronger effect should have been easier to detect in the RNA-seq data.
  - How strong was the uncorrected p-value? A more significant eQTL should have been more reproducible.

For each dichotomous (boolean) variable, I counted the number of concordant cases for each of the groups. I tested how significantly different these numbers were using a two-tailed Fisher's exact test. Two examples of the fractions of concordant cases for different variables are shown in Figure 5.5.



(a) Is the variant a SNP or indel?

(b) Is the variant within 25 bp of a peak?

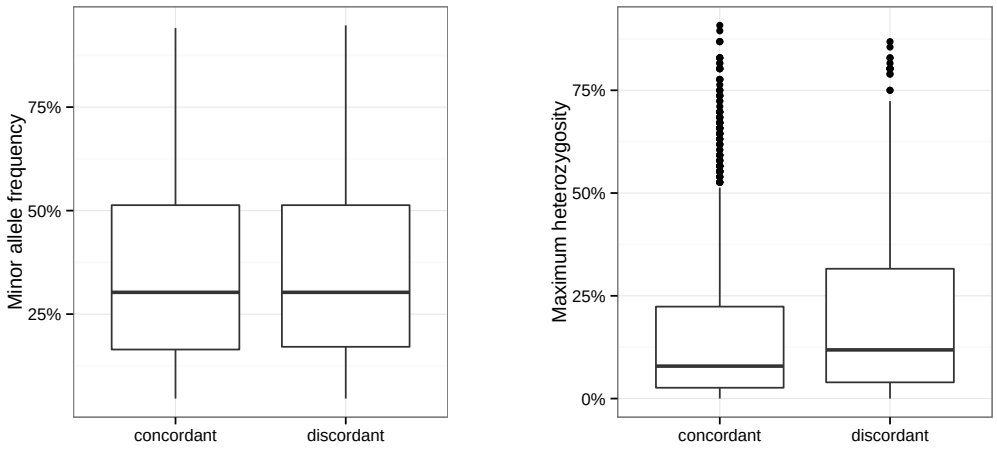
Figure 5.5.: Fractions of concordant lead gene eQTLs for two example boolean variables.

The first variable (Figure 5.5a) shows whether the associated variant was an indel or not. This variable did not seem to have any influence on the concordance rate, as shown by the uniform plot and the resulting p-value of 1. In contrast, the variable indicating whether the variant was within 25 bp of the peak (Figure 5.5b) did seem to have a large effect on the concordance rate, with $p = 1.90 \cdot 10^{-8}$.

Similarly, for each continuous variable, I calculated the difference in value between concordant and discordant cases using a Wilcoxon rank sum test. Two examples of such variables are shown in Figure 5.6.

The first variable, the minor allele frequency of the variant (Figure 5.6a), did not appear to be associated with the concordance rate, with $p = 0.72$. However, the variable that reflected the maximum heterozygosity rate of variants in the peak regions (Figure 5.6b) showed a clear association with the concordance rate ($p = 1.97 \cdot 10^{-8}$), with higher values being associated with lower concordance.

From the set of continuous variables, I selected two variables for further optimisation, which were strongly associated with the concordance rate. These were the maximum heterozygosity rate ($p = 1.90 \cdot 10^{-8}$) and the p-value of the correlation between the genotypes and the mappability of the peak regions ($p = 1.84 \cdot 10^{-3}$).



(a) MAF of the variant **(b)** Maximum degree of heterozygosity

Figure 5.6.: Fractions of concordant lead eQTLs for two example continuous variables.

The fact that both of these correlated with the concordance rate again agreed with the expectation that a large fraction of the discordant eQTL calls might have been false positives caused by mappability problems, especially in heterozygous regions.

For each of them, I chose several possible threshold values and then calculated the concordance rate as well as the fraction of eQTLs passing the filters. The resulting data points are shown in Figure 5.7.

Based on these results I set the thresholds for maximum heterozygosity to 40 % and the threshold for the p-value of the mappability correlation to 10^{-6} as these appeared to represent a good compromise between increase in concordance and loss of eQTLs. The data point showing the expected results of these thresholds is highlighted as a triangle in Figure 5.7. For the common gene eQTLs, this resulted in a set of 2,095 genes with eQTLs, down from 3,139 genes in the unfiltered set (67 %).

Additionally, I noted that the boolean variable indicating whether the variant was close to a peak or not appeared to have a large effect on the concordance rate ($p = 1.90 \cdot 10^{-8}$). This suggested that there were additional problems with variants close to 3′ Tag-Seq regions, which were not detected by the mappability filters. I thus removed all genes with their lead variant within 25 bp of the 3′ Tag-Seq peak, which resulted in a further reduction to 1,596 cases 51 %.

This filter, by definition, removed almost all signal close to the 3′ end of genes. However, my analysis of high-confidence eQTLs had shown that there seemed to

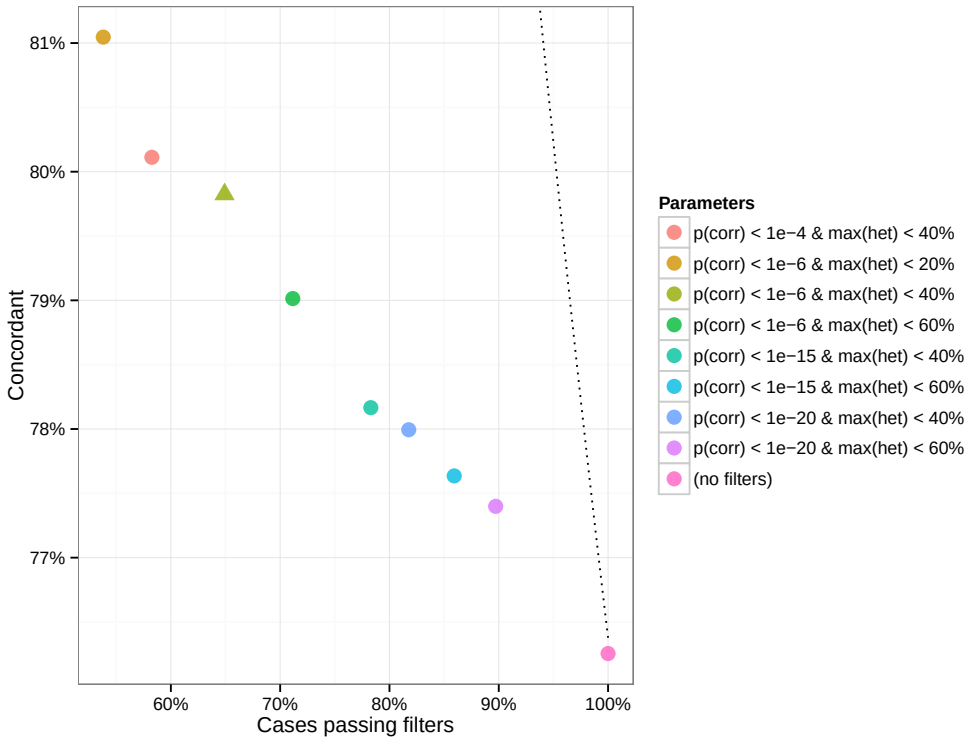


Figure 5.7.: Percentage of concordant eQTLs plotted against the percentage of eQTLs passing the filters. Coloured points show eight different combinations of filtering thresholds on the p-value of the correlation between the genotypes and the mappability of the peak regions and the maximum heterozygosity rate. Chosen thresholds shown as triangle. Dotted line shows the results that could be obtained with an optimal filter, under the conservative assumption that the RNA-seq estimates always describe the true expression level.

be an enrichment of real eQTLs at the 3′ end even in the RNA-seq data, which suggested that this filter might have removed a substantial amount of real signal as well. Thus, to account for this highly conservative filter, I also generated a second set of eQTLs without this filter, which I used to study specifically cases of eQTLs at the 3′ end of genes.

Finally, I also removed all genes for which I had found a very large number of associated variants, or variants very far apart from each other, i.e. where the eQTL cloud was large. This was because these were likely to be regions of strong LD where it would be unlikely that I had identified the real causal variant as my lead eQTL. I thus removed all genes where the gene had more than 10 variants in its eQTL cloud or where the variants in the eQTL cloud spanned more than 10 kb.

For the common gene eQTLs this resulted in the removal of a further 372/1,596 of genes and a final eQTL set of 1,224 genes (39 % of the unfiltered set).

It should be noted that this filter will also have removed cases where there are actually two independent genetic effects acting on the gene. However, due to the relatively low sample size I would have been underpowered to systematically detect cases like these, as discussed in Section 4.6.1. Visual inspection of some of the genes removed by the cloud filter also did not indicate the presence of a large amount of such cases.

After applying all of these filters, I observed a final concordance rate of 83 % in the set of lead gene eQTLs with a common effect (Figure 5.8).

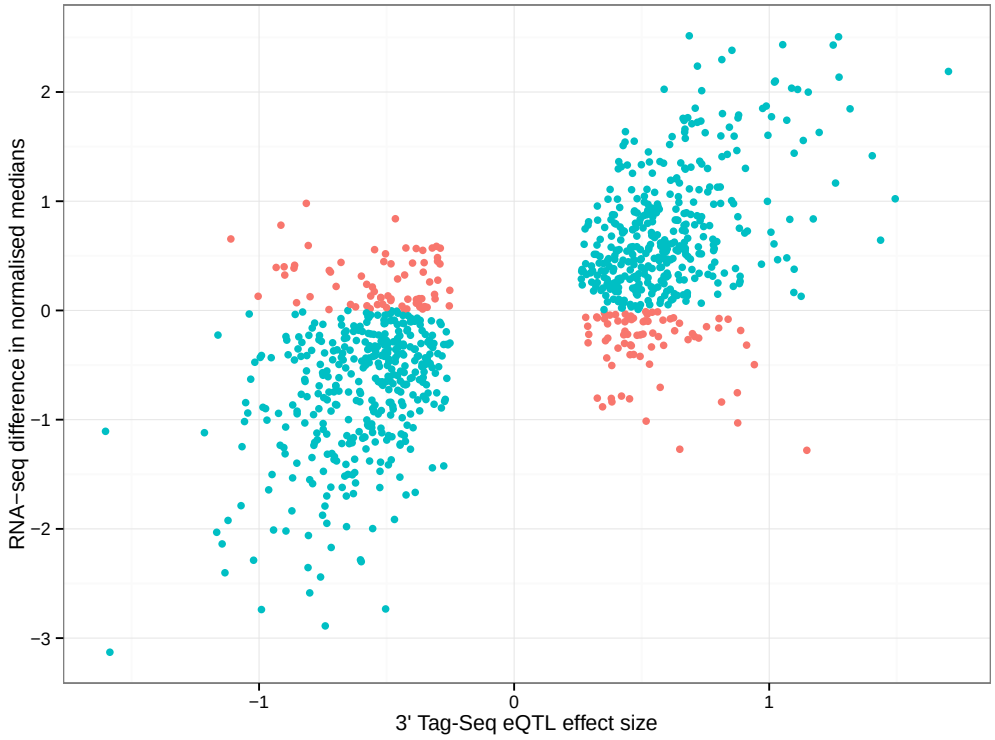


Figure 5.8.: Difference in normalised medians for expression levels measured with standard RNA-seq at 10–12 h plotted against effect size at 10–12 h of 3′ Tag-Seq eQTLs. Concordant eQTLs shown in blue, discordant eQTLs shown in red.

Having found this set of filters to be appropriate for decreasing the false positive rate caused by protocol-specific effects, I applied them to all eQTL sets. This included both the set of stage-specific lead eQTLs, as well as the four sets of eQTL clouds, where I removed the entire eQTL cloud if the lead variant did not

pass the filters and also removed individual variants from eQTL clouds if they did not pass the filters themselves.

Finally, I applied all but the close-to-peak filter to all peak eQTL sets as well. Out of 10,387 common peak lead eQTLs, I was left with 7,296 eQTLs after the mappability filters and 5,883 eQTLs after the cloud filter (57 %).

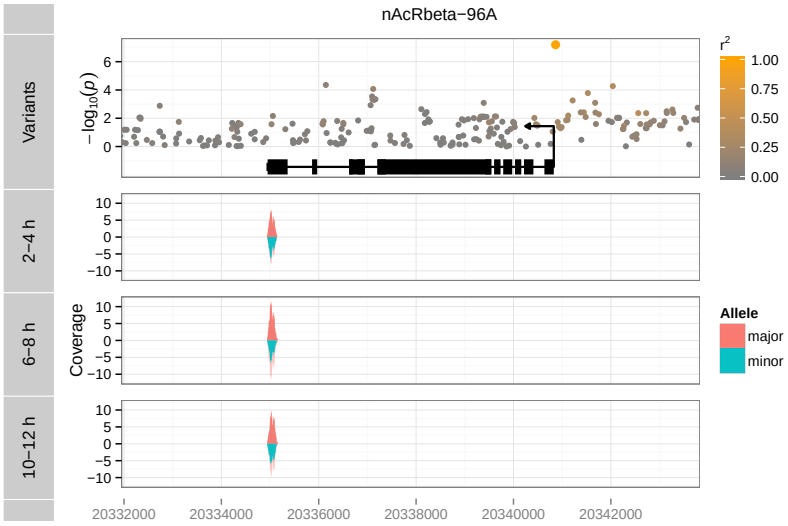
5.3. Examples of eQTLs

Before I analysed these filtered sets of eQTLs in aggregate (see Chapter 6) I inspected a few individual genes to get an impression of what these eQTLs look like in detail.

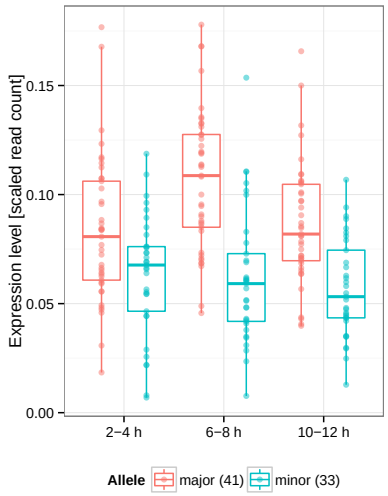
An example for an common eQTL associated with the gene *nicotinic Acetylcholine Receptor β2* (*nAChRβ2*) is shown in Figure 5.9. *nAChRβ2* is a subunit of the nicotinic Acetylcholine receptor, which is involved in the central nervous system of *Drosophila* (Gundelfinger and Hess, 1992). I observed a single 3′ Tag-Seq peak for this gene, consistent with the single 3′ end isoform that has been observed previously (The FlyBase Consortium, 2014). The lead eQTL, a T to C SNP, was located 37 bp upstream of the 5′ end of *nAChRβ2* and had a MAF of 46 % in the DGRP. This nucleotide change was negatively associated with the gene's expression level, with individuals homozygous for the major allele showing almost twice the level of expression as individuals homozygous for the minor allele at 6–8 h.

The top of the first panel (Figure 5.9a) shows the body of the gene (in black), with the boxes indicating exons and the arrow indicating the orientation of the gene, as well as all tested variants around it (circles). For each tested variant, its Y-position shows the strength of association of that variant with the expression level of the gene, as expressed by the $-log_{10}(p)$ value of the common-effect test. This manner of depicting the association of variants by their genomic location is called a Manhattan plot. The lead variant is shown as a yellow circle, and all other variants are coloured by their r^2 to this variant, indicating the strength of LD between them.

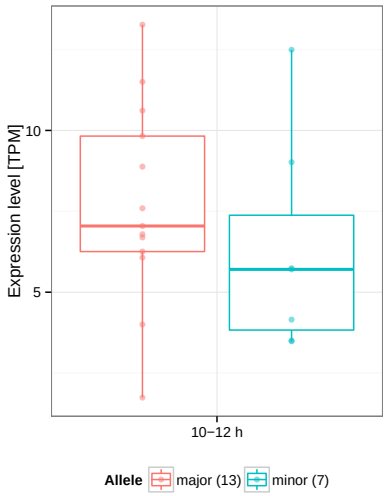
Below the variants and the gene body, the 3′ Tag-Seq coverage associated with this gene at the three developmental stages is shown. This shows the median read coverage of all lines with the major allele (red, above 0) and the minor allele (blue, below 0) at the lead eQTL. For ease of comparison, the coverage is reflected on the X-axis in a lighter colour. For readability only the read coverage inside the 3′



(a) Manhattan plot and median 3′ Tag-Seq coverage



(b) 3′ Tag-Seq expression level



(c) RNA-seq expression level

Figure 5.9.: eQTL for the gene *nAChRβ2*. (a) Top: Manhattan plot of variants around the gene body (black). Below: Median 3′ Tag-Seq coverage at 2–4 h, 6–8 h and 10–12 h, grouped by major/minor allele at the lead eQTL. Only reads associated with *nAChRβ2* shown. Bottom: Boxplots of 3′ Tag-Seq (b) and RNA-seq (c) expression levels. Samples with heterozygous genotype at the lead eQTL omitted.

Tag-Seq peaks associated with this gene is shown.

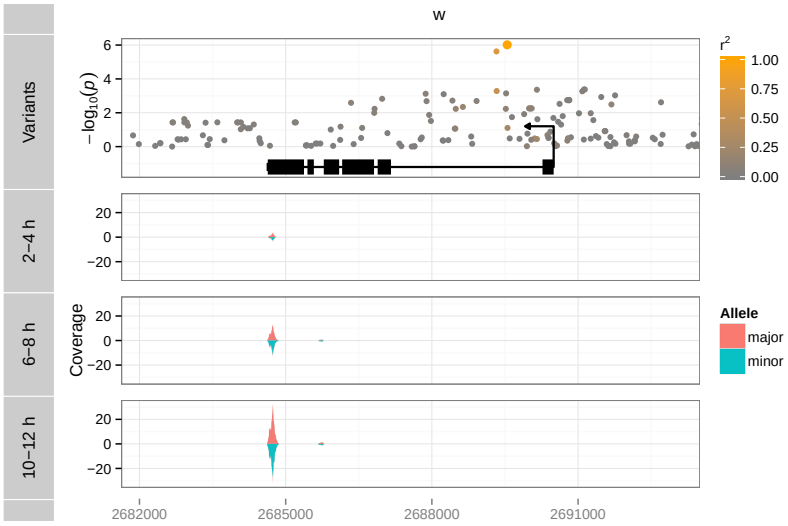
The bottom left panel (Figure 5.9b) shows a comparison of the gene expression levels at each of the three developmental stages between the major (red) and the minor (blue) allele, as estimated from scaled 3′ Tag-Seq read counts. The boxplots indicate the overall distribution of the expression levels, while the points

show each individual line's expression level. The same comparison is repeated based on the RNA-seq data for 10–12 h in the bottom right panel (Figure 5.9c).

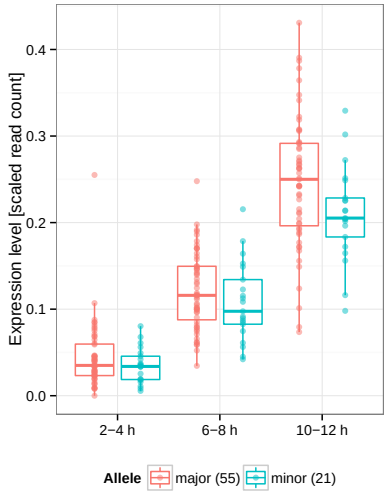
I also found eQTLs for two classic *Drosophila* genes described in Section 1.1: *white* and *Dichaete* (see Section 1.1.2). The eQTL plots for these genes are shown in Figures 5.10 and 5.11.

The lead eQTL for *white* was located it its first intron, approximately 1 kb downstream from the gene's 5′ end. It had a MAF of 23 % and resulted in a change from an A to a T nucleotide, which was associated with a weak negative effect on the expression level. As before, I only observed a single 3′ Tag-Seq peak for this gene, consistent with the existing annotation.

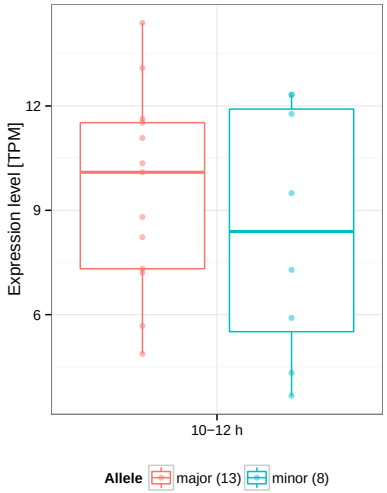
For *Dichaete*, I observed two separate 3′ Tag-Seq peaks, matching the location of the two alternate poly(A) sites that are known for this gene. The lead eQTL was located very close to the secondary, less highly expressed, 3′ Tag-Seq peak, 15 bp downstream of the annotated cleavage site. This would have normally resulted in a removal of this gene by the filters, but I retained the eQTL for this analysis after closer inspection did not reveal any likely mappability problems and the RNA-seq data was concordant. The lead variant was an A to C SNP with a MAF of 9 %. It was associated with a decreased expression level of the secondary 3′ isoform, which consequently decreased the overall expression level of the gene.



(a) Manhattan plot and median 3′ Tag-Seq coverage

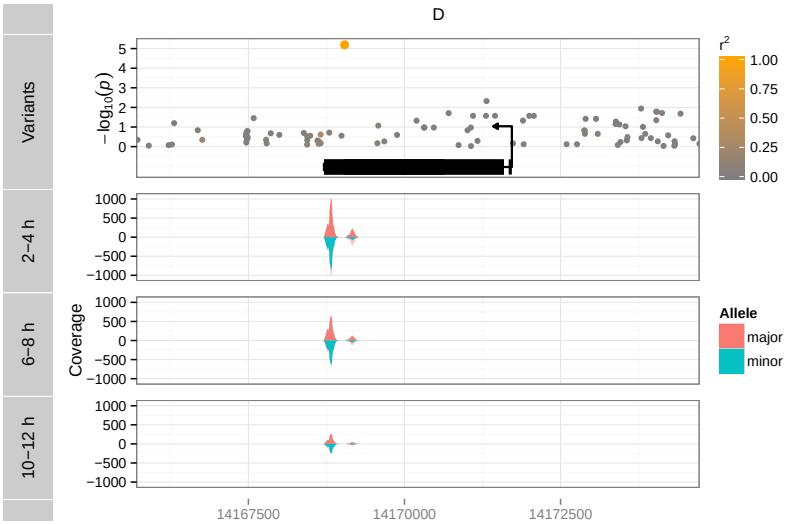


(b) 3′ Tag-Seq expression level

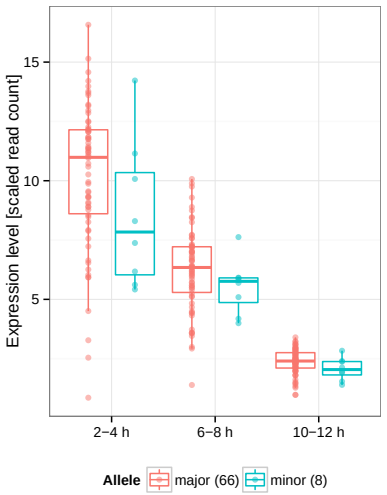


(c) RNA-seq expression level

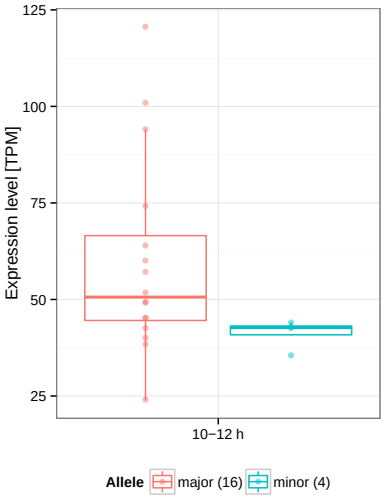
Figure 5.10.: eQTL for the gene *white*. (a) Top: Manhattan plot of variants around the gene body (black). Below: Median 3′ Tag-Seq coverage at 2–4 h, 6–8 h and 10–12 h, grouped by major/minor allele at the lead eQTL. Only reads associated with *white* shown. Bottom: Boxplots of 3′ Tag-Seq (b) and RNA-seq (c) expression levels. Samples with heterozygous genotype at the lead eQTL omitted.



(a) Manhattan plot and median 3′ Tag-Seq coverage



(b) 3′ Tag-Seq expression level



(c) RNA-seq expression level

Figure 5.11.: eQTL for the gene *Dichaete*. (a) Top: Manhattan plot of variants around the gene body (black). Below: Median 3′ Tag-Seq coverage at 2–4 h, 6–8 h and 10–12 h, grouped by major/minor allele at the lead eQTL. Only reads associated with *Dichaete* shown. Bottom: Boxplots of 3′ Tag-Seq (b) and RNA-seq (c) expression levels. Samples with heterozygous genotype at the lead eQTL omitted.

5.4. Validation of eQTLs by *in situ* hybridisation

As a final quality control step for my eQTLs, I also wanted to validate the impact of some of the eQTLs I had found using an *in vivo* experiment. To this end, my collaborator Enrico Cannavò (a PhD student from the Furlong group in EMBL-Heidelberg) and I selected a set of genes with eQTLs, on which Enrico then performed RNA *in situ* hybridisation (Levsky and Singer, 2003). By comparing the *in situ* hybridisation results from flies between the major and minor allele, we could test whether the eQTL did in fact result in a change in the amount of RNA present in the embryo. Using this method, we could successfully validate the eQTL for the gene *GstD1*, for which I had observed a very large change in gene expression between the two alleles. As predicted by the eQTL, expression of *GstD1* was only observed in the fly with the major allele and was also consistent between 6–8 h and 10–12 h, as shown in Figure 5.12). An eQTL plot for this gene is shown in Figure 5.13.

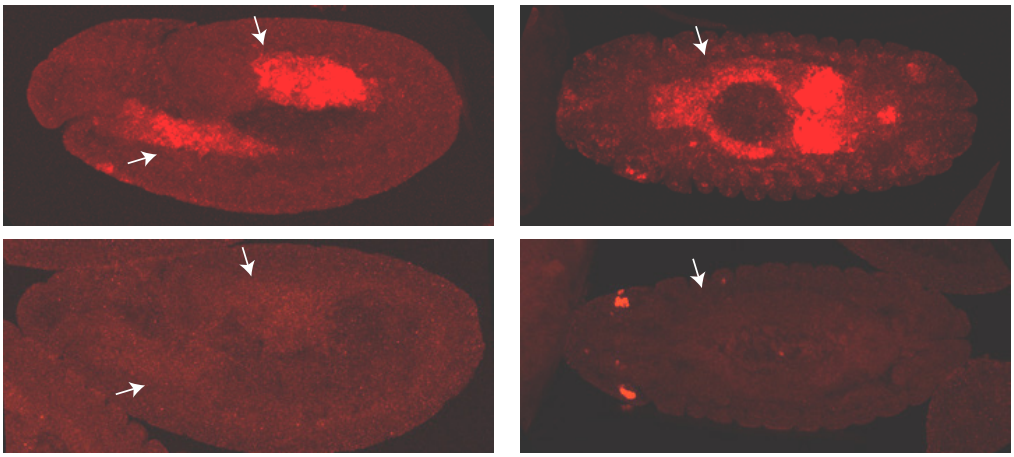
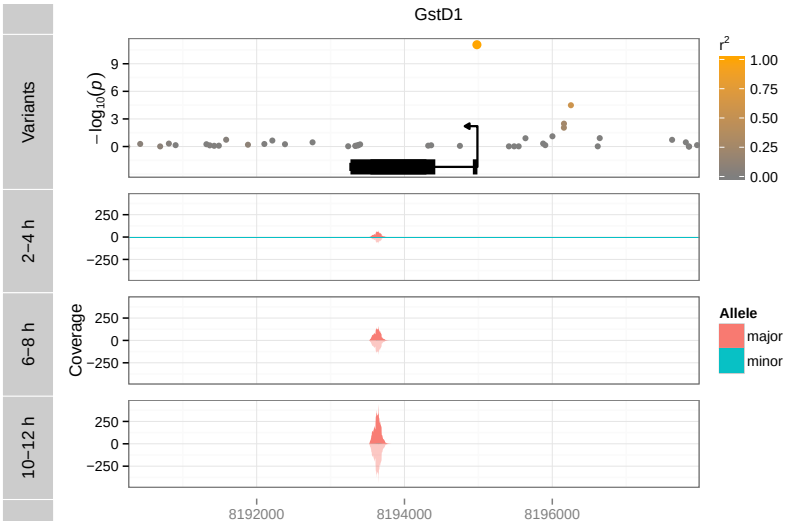
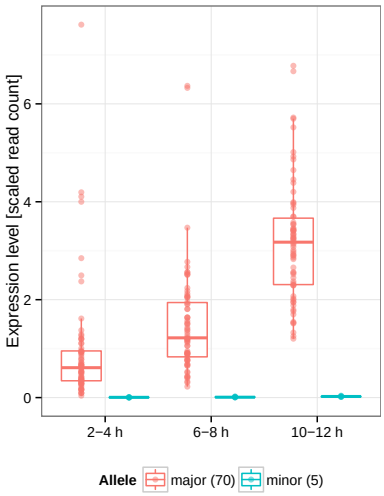


Figure 5.12.: *In situ* hybridisation of *GstD1* (red) in embryos homozygous for the major (top) and minor (bottom) allele. Collected at 6–8 h (left) and 10–12 h (right) after fertilisation. White arrows indicate the midgut, which shows high levels of *GstD1* only in the major allele. Microscopy and figure by Enrico Cannavò.

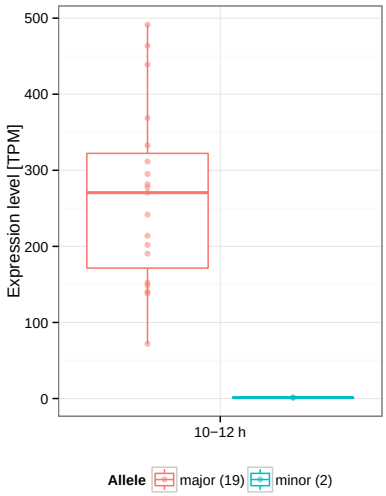
The *in situ* hybridisation experiment showed that variants that I found to be associated with the 3′ Tag-Seq expression levels indeed resulted in molecular changes in the organism and could be validated even with this completely different technology. In addition, Enrico Cannavò also performed a further set of experiments on another three genes with a single, promoter-proximal, strongly associated SNP as



(a) Manhattan plot and median 3′ Tag-Seq coverage



(b) 3′ Tag-Seq expression level



(c) RNA-seq expression level

Figure 5.13.: eQTL for the gene *GstD1*. (a) Top: Manhattan plot of variants around the gene body (black). Below: Median 3′ Tag-Seq coverage at 2–4 h, 6–8 h and 10–12 h, grouped by major/minor allele at the lead eQTL. Only reads associated with *GstD1* shown. Bottom: Boxplots of 3′ Tag-Seq (b) and RNA-seq (c) expression levels. Samples with heterozygous genotype at the lead eQTL omitted.

their eQTL. For each eQTL, he determined the DNA sequence around that SNP, both for strains with the major allele and for strains with the minor allele. Using an *in vitro* Luciferase reporter assay, he could then test whether these stretches of DNA resulted in a change in expression of the adjacent gene. In all three cases, he was able to recapitulate the change predicted by the eQTL. These results will

be reported in more detail in our joint paper (Cannavo et al., 2015).

5.5. Comparison of eQTLs with a previously published study

In 2012, Massouras and colleagues published the first eQTL study in the DGRP (Massouras et al., 2012). Their study differed from our study in multiple ways, namely:

- They tested samples from a single stage (adult flies), not from multiple developmental stages.
- They obtained gene expression levels for 39 lines, resulting in low statistical power.
- They estimated gene expression levels from microarrays instead of from a sequencing-based method.
- They did not perform any mappability filtering.

Due to these major differences, a large amount of overlap between my sets of eQTLs and the Massouras set was not likely. In particular, the lack of mappability filtering and their reliance on less accurate expression level estimates (microarrays) may have led to an increased amount of false positives in their study. Nevertheless, I would expect that genes with an eQTL in my study would have an increased chance of also having an eQTL in the Massouras study.

To test this assumption, I overlapped the set of genes for which I had found a significant eQTL in my study with the set of genes with eQTLs from Massouras and colleagues. Out of the 1,224 genes that had a significant common eQTL after filtering, 295 (24 %) also had an eQTL in the older study. Out of the 8,870 genes that I had tested but that did not have an eQTL in my study, 1,333 (15 %) had an eQTL in theirs. This difference constitutes a significant enrichment (Fisher's exact test, $p = 1.12 \cdot 10^{-14}$).

In fact, I identified exactly the same variant as the eQTL for 73 out of the 295 genes with overlaps. For 69 of these (98 %) the direction of the predicted effect was also the same in both studies.

This shows that, while there are clear differences between the eQTL sets, there is still considerable overlap, which speaks to the true signal present in both of these completely independently obtained eQTL sets.

6. Analysis of gene-proximal eQTLs

In Chapter 4, I described how I used a multivariate linear mixed model to jointly call eQTLs in the three developmental stages in this study: 2–4 h, 6–8 h and 10–12 h after fertilisation. I tested each gene and each 3′ isoform for association with variants in a ±50 kb window around the gene. After applying a set of filters to remove possible false positives in Chapter 5, I obtained four sets of eQTLs each for genes and 3′ isoforms: a set of eQTLs common to all three developmental stages, and three sets of eQTLs with specific effects in one of the developmental stages. In this chapter, I describe the properties of these eQTL sets.

6.1. Properties of common gene eQTLs

For the first part of my analysis, I concentrated on the set of 1,224 lead gene eQTLs with a common effect, since they made up the largest part of the eQTLs and thus provided me with the greatest power to detect interesting patterns.

6.1.1. Enriched and depleted gene categories

A naive view of development would be that any changes to the timing or level of the expression of developmental genes would result in abnormal development or death of the embryo before it reaches adulthood. On the other hand, genes that are not important for development would not be under such strong constraints and should thus be more likely to tolerate changes in expression level. In line with this, earlier studies of gene expression level variation in *Drosophila* have shown that the expression levels of genes important in development are more constrained than those of other genes (Rifkin et al., 2003; Rifkin et al., 2005; Kalinka et al., 2010).

Consequently, I would expect that I would be less likely to find eQTLs for genes associated with development. In order to test this hypothesis, I performed a GO enrichment/depletion analysis comparing the set of genes with an eQTL to the full set of genes that I had tested.

In order to account for possible confounding effects of the filters I applied to the eQTLs (see Chapter 5), I applied the same filters to the background set of tested genes whenever possible. In particular, I removed all tested genes where the most strongly associated (not necessarily significant) variant was close to a 3′ Tag-Seq peak or was associated with mappability or where there was a variant with high heterozygosity in one of the gene's peaks.

After accounting for these factors, I found 6 GO categories that were significantly enriched and 69 that were significantly depleted for genes with common eQTLs, using a two-tailed Fisher's exact test with topGO. I only tested GO categories with a minimum size of 100 annotated genes and used the BH method to control the FDR. A selection of significantly depleted and enriched GO terms in the "biological process" (BP) and "molecular function" (MF) gene ontologies at an FDR of 5 % are shown in Tables 6.1 and 6.2. There were no enriched or depleted terms in the "cellular component" (CC) ontology at this significance threshold.

Term	Exp.	Obs.	Dir.	FDR
carboxylic acid metabolic process	33.00	49	↑ enr.	0.02
organic acid metabolic process	34.66	50	↑ enr.	0.03
oxoacid metabolic process	34.66	50	↑ enr.	0.03
cellular amino acid metabolic process	24.34	37	↑ enr.	0.04
single-organism metabolic process	136.62	162	↑ enr.	0.05
organ development	117.81	70	↓ depl.	0.00
single-organism developmental process	201.70	151	↓ depl.	0.00
organ morphogenesis	59.18	33	↓ depl.	0.00
single-organism process	513.83	460	↓ depl.	0.00
regulation of multicellular organismal p...	45.54	23	↓ depl.	0.00
anatomical structure development	254.61	208	↓ depl.	0.00
regulation of cellular process	243.37	197	↓ depl.	0.00
regulation of biological process	262.17	216	↓ depl.	0.01
imaginal disc development	53.28	30	↓ depl.	0.01
regulation of developmental process	42.22	22	↓ depl.	0.01

Table 6.1.: Terms in the "biological process" ontology significantly enriched (enr.) or depleted (depl.) for eQTLs at an FDR threshold of 5 %. Only top 10 out of 62 depleted terms shown. Exp., expected number of genes. Obs., observed number of genes. Dir., direction.

As expected, there was a strong depletion of genes with eQTLs in many developmentally important categories, such as "regulation of developmental process",

Term	Exp.	Obs.	Dir.	FDR
catalytic activity	382.63	426	↑ enr.	0.02
sequence-specific DNA binding	31.03	13	↓ depl.	0.01
nucleic acid binding transcription facto...	46.27	27	↓ depl.	0.02
DNA binding	76.75	53	↓ depl.	0.02
sequence-specific DNA binding transcript...	46.27	27	↓ depl.	0.02
binding	492.09	450	↓ depl.	0.02
protein binding	136.96	108	↓ depl.	0.03
structural molecule activity	50.18	33	↓ depl.	0.04

Table 6.2.: Terms in the "molecular function" ontology significantly enriched (enr.) or depleted (depl.) for eQTLs at an FDR threshold of 5 %. Exp., expected number of genes. Obs., observed number of genes. Dir., direction.

"organ development", and "imaginal disc development". Similarly, the molecular functions depleted in eQTLs included terms such as "DNA binding" and "protein binding" and "structural molecule activity".

On the other hand, I observed an enrichment of GO categories related to metabolic processes and catalytic activity. This agrees with earlier work which had suggested that the developmental expression levels of metabolic genes are under less selection and are thus more free to vary between individuals (Kalinka et al., 2010).

6.1.2. eQTLs in developmentally important genes

While there was a general depletion in developmental genes among those with eQTLs, there were some exceptions — for example, I had found eQTLs for 32 genes annotated with the GO term "embryo development". These eQTLs were of particular interest, as they provided a set of developmental genes for which gene expression levels appeared to diverge between individuals due to genetic variation. The variants associated with these loci reached minor allele frequencies of up to 49 %, suggesting that their change in expression level was not under negative selection.

One possible reason for this would be that these genes were not actually important in the developmental stage in which they were affected by the eQTLs. One could thus argue that these eQTLs just represented fluctuations of no biological consequence and that the genes were much more tightly regulated in the stages where they were actually important. To account for this concern, I looked for embryo development genes that were strongly expressed during one of our devel-

opmental stages, suggesting that they had a function in that stage. Using the modENCODE developmental time course data (see Chapter 3), I calculated the maximum expression level during embryo development for each gene and then compared the expression level at each of my three developmental stages to this maximum. I considered each gene that was expressed to at least 70 % of its maximum expression level at one of the three developmental stages to be strongly expressed. For each of the strongly expressed genes, I also checked whether I had observed a concordant significant difference in the RNA-seq data, as described in Chapter 5.

I found 22 genes with common gene eQTLs that were annotated with "embryo development" and strongly expressed in at least one of the time points in my study. A list of these 22 genes and the properties of their eQTLs is shown in Table 6.3.

For three of these 22 genes I had also observed a significant effect in the RNA-seq data (Wilcoxon test, unadjusted $p < 0.05$), suggesting that their eQTLs were particularly strong and reliable. These were *brn* (*brainiac*), *ecd* (*ecdysoneless*) and *Orct2* (*Organic cation transporter 2*). Further literature review revealed that these genes indeed have known functions during embryo development:

- *brainiac*: Most highly expressed at 2–4 h, $\log_2$(fold-change) $= -0.18$ at this stage (decrease). Multiple associated variants in the only exon.

  This gene is a Glycolipid-specific β1,3N-Acetylglucosaminyltransferase (Müller et al., 2002), which is known to be important for epithelial morphogenesis during *D. melanogaster* oogenesis (Goode et al., 1996).

- *ecdysoneless*: Most highly expressed at 6–8 h, $\log_2$(fold-change) $= 0.71$ at this stage (increase). Multiple associated variants in the second (and last) exon. While this gene was most strongly expressed at 6–8 h in individuals with the reference allele, its expression was actually higher at 2–4 h for individuals with the alternate allele.

  This gene is associated with the production of the steroid hormone ecdysone which is important for the coordination of many developmental processes, including embryogenesis (Gaziova et al., 2004). Disruption of ecdysone production by a mutation in this gene has been associated with a variety of developmental defects.

- *Organic cation transporter 2* (also known as *calderón*): Most highly ex-

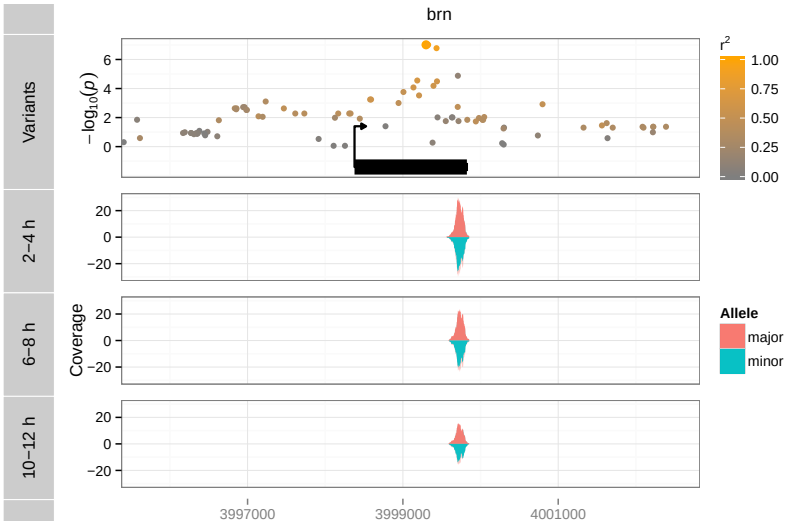
			2–4 h		6–8 h		10–12 h	
Gene	Lead variant	MAF	Exp.	$\log_2$(fc)	Exp.	$\log_2$(fc)	Exp.	$\log_2$(fc)
aub	2L:11033738	0.04	0.87	1.62	0.42	1.35	0.24	1.55
ecd *	3L:2264533	0.06	0.57	1.13	0.57	0.71	0.37	0.72
jeb	2R:7949532	0.06	0.14	0.93	0.28	0.31	0.16	0.15
eIF4AIII	3R:4112706	0.04	5.01	0.89	3.30	0.47	2.03	0.26
phm	X:18580422	0.09	0.84	-0.84	0.06	-0.87	0.07	-0.68
Sec61α	2L:6496440	0.04	2.40	0.84	6.85	0.55	7.93	0.25
htl	3R:13922002	0.04	0.71	-0.39	0.68	-0.75	0.77	-0.54
csul	2R:12122214	0.05	1.08	-0.72	0.57	-0.26	0.30	-0.08
faf	3R:27611336	0.26	2.37	0.70	1.36	0.40	1.85	0.22
Orct2 *	3R:20098687	0.30	0.36	0.55	1.85	0.34	0.37	0.38
Rop	3L:4136262	0.49	0.90	0.28	1.19	0.48	1.93	0.34
aPKC	2R:10834039	0.24	2.09	-0.48	1.92	-0.24	1.29	-0.23
nst	3L:12526237	0.46	1.46	0.46	1.11	0.31	1.63	0.30
stumps	3R:10423593	0.34	2.03	-0.41	2.65	-0.19	2.13	-0.05
dia	2L:20726434	0.13	0.83	-0.35	0.84	-0.36	0.69	-0.34
Nc	3L:9964959	0.33	2.18	0.12	1.47	0.32	1.37	0.19
brn *	X:3999298	0.31	0.26	-0.18	0.20	-0.10	0.13	-0.31
numb	2L:9448470	0.07	1.97	-0.11	1.94	-0.26	1.85	-0.30
inx2	X:6943533	0.18	21.76	0.30	35.85	0.19	24.93	0.17
Hakai	2L:19598306	0.09	1.23	-0.28	1.08	-0.07	0.59	-0.25
Mad	2L:3147509	0.09	1.83	0.27	1.90	0.18	1.66	0.26
sqh	X:6121036	0.19	9.53	0.20	11.83	0.07	10.11	0.11

Table 6.3.: Lead common eQTLs associated with strongly expressed genes that were annotated with the GO term "embryo development" (GO:0009790) or its children. Exp., mean expression level. $\log_2$(fc), $\log_2$ of fold-change between the minor and major allele. *, significant effect also observed in RNA-seq data. Sorted by absolute $\log_2$(fold-change).

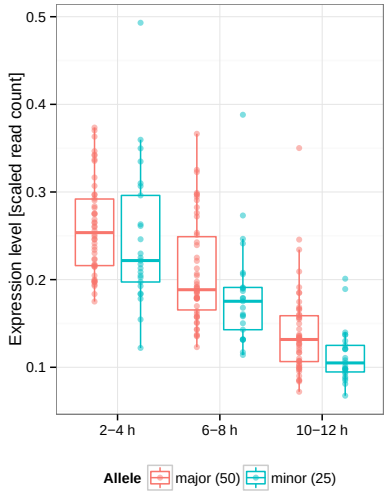
pressed at 6–8 h, $\log_2$(fold-change) = 0.34 at this stage (increase). Single strong lead variant in the promoter region.

Loss of this gene has been associated with an inability of embryos to retract the germ band during embryogenesis, a delay in developmental speed and a decrease in the body size of adult flies (Herranz et al., 2006).

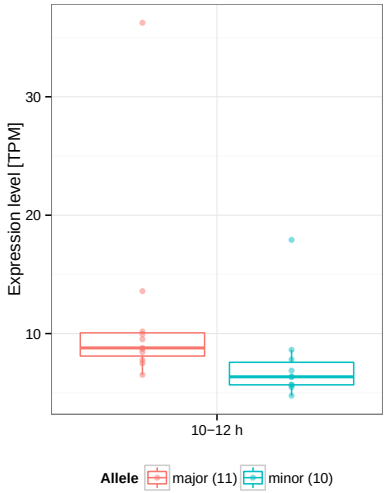
Manhattan plots for the three genes, together with the median 3′ Tag-Seq coverage and the boxplots of 3′ Tag-Seq and RNA-seq expression levels compared between the two genotypes are shown in Figures 6.1 to 6.3.



(a) Manhattan plot and median 3′ Tag-Seq coverage

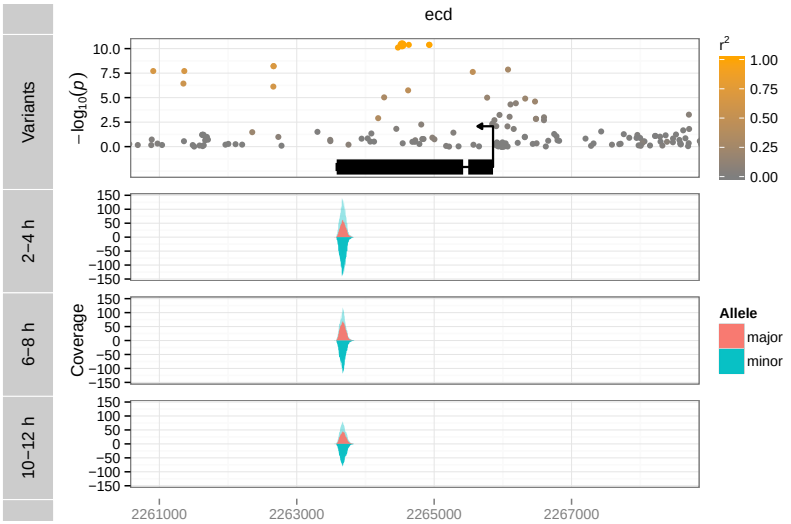


(b) 3′ Tag-Seq expression level

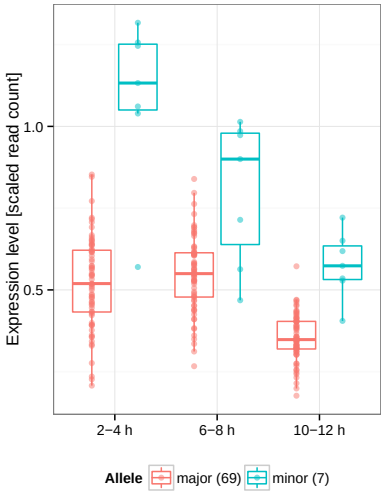


(c) RNA-seq expression level

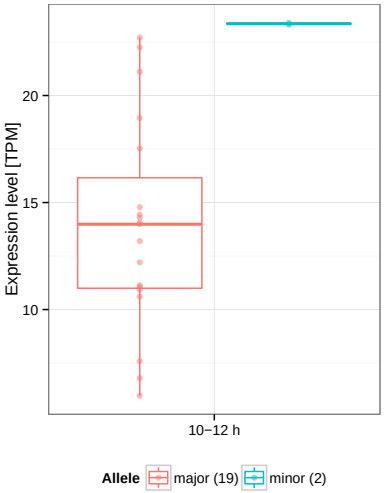
Figure 6.1.: eQTL for the gene *brn*. (a) Top: Manhattan plot of variants around the gene body (black). Below: Median 3′ Tag-Seq coverage at 2–4 h, 6–8 h and 10–12 h, grouped by major/minor allele at the lead eQTL. Only reads associated with *brn* shown. Bottom: Boxplots of 3′ Tag-Seq (b) and RNA-seq (c) expression levels. Samples with heterozygous genotype at the lead eQTL omitted.



(a) Manhattan plot and median 3′ Tag-Seq coverage

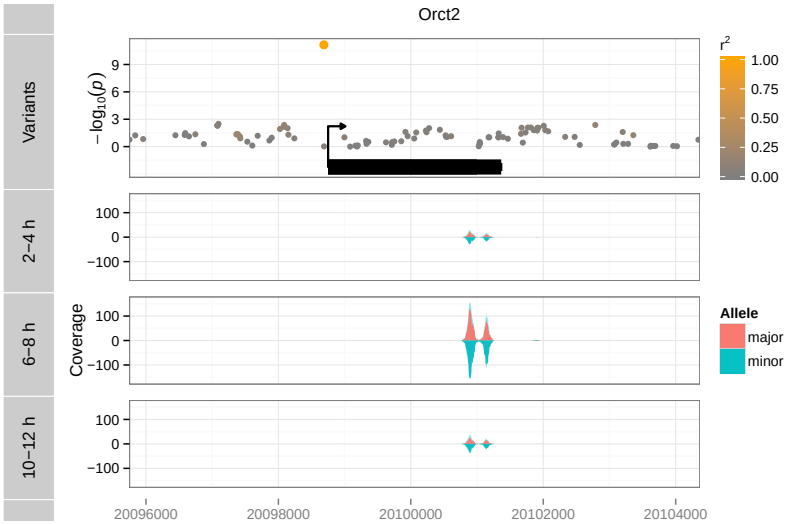


(b) 3′ Tag-Seq expression level

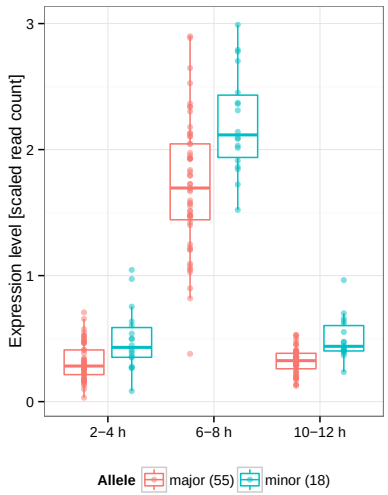


(c) RNA-seq expression level

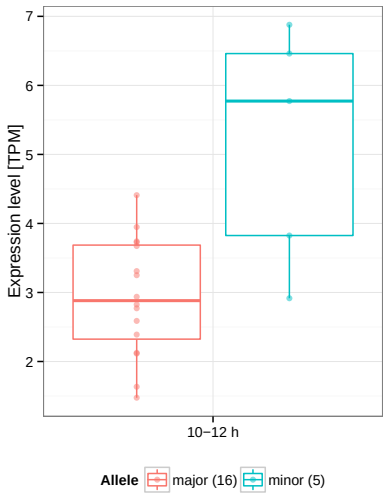
Figure 6.2.: eQTL for the gene *ecd*. (a) Top: Manhattan plot of variants around the gene body (black). Below: Median 3′ Tag-Seq coverage at 2–4 h, 6–8 h and 10–12 h, grouped by major/minor allele at the lead eQTL. Only reads associated with *ecd* shown. Bottom: Boxplots of 3′ Tag-Seq (b) and RNA-seq (c) expression levels. Samples with heterozygous genotype at the lead eQTL omitted.



(a) Manhattan plot and median 3′ Tag-Seq coverage



(b) 3′ Tag-Seq expression level



(c) RNA-seq expression level

Figure 6.3.: eQTL for the gene *Orct2*. (a) Top: Manhattan plot of variants around the gene body (black). Below: Median 3′ Tag-Seq coverage at 2–4 h, 6–8 h and 10–12 h, grouped by major/minor allele at the lead eQTL. Only reads associated with *Orct2* shown. Bottom: Boxplots of 3′ Tag-Seq (b) and RNA-seq (c) expression levels. Samples with heterozygous genotype at the lead eQTL omitted.

Since a loss of *Orct2* had previously been associated with a decrease in the body size of flies, I tested whether there was an association between its eQTL and a recently published set of adult body size phenotypes (Vonesch et al., 2015). In their study, Vonesch and colleagues phenotyped adults from 149 DGRP lines for a measure of size based on their wing imaginal disc (the centroid size) and one based on their eye imaginal disc (interocular distance). Interestingly, the centroid size of male adult flies was indeed slightly increased for flies with the minor genotype at the SNP (one-sided Wilcoxon rank-sum test, $p = 0.09$), in line with my observation that expression of *Orct2* was higher for flies with the minor genotype. However, accounting for the fact that I tested four phenotypes for association (male/female centroid size, male/female interocular distance) the Bonferroni-corrected p-value accounting for multiple testing would have only been 0.36.

To explore this further, I plotted the male centroid size against the expression level of *Orct2* and then coloured each line by its genotype at the eQTL. Using this approach, I could test whether there was a correlation between *Orct2* expression and size even without the eQTL or whether the correlation was entirely driven by the genotype. If the former were the case, there would be a correlation between expression level and size within each genotype as well, while if the latter were the case there would be no clear correlation. The result of this analysis is shown in Figure 6.4.

As expected, given my small sample size and the complexity of this whole-body phenotype, I did not have enough data to reach any confident conclusions about the association between gene expression levels and body size. Nevertheless, the data seems to suggest that while the size estimate is overall positively associated with the gene expression level (black line, Spearman's $\rho = 0.16$), this is mainly driven by a strong positive association in flies homozygous for the minor allele (blue line, Spearman's $\rho = 0.43$). For flies homozygous for the major allele, there actually seems to be a weak negative correlation (red line, Spearman's $\rho = -0.11$). This interaction is suggestive of a complex relationship between this SNP, the gene expression of *Orct2* and body size. While not statistically significant, this is certainly an interesting case which might warrant further study, as I will discuss in Chapter 8.

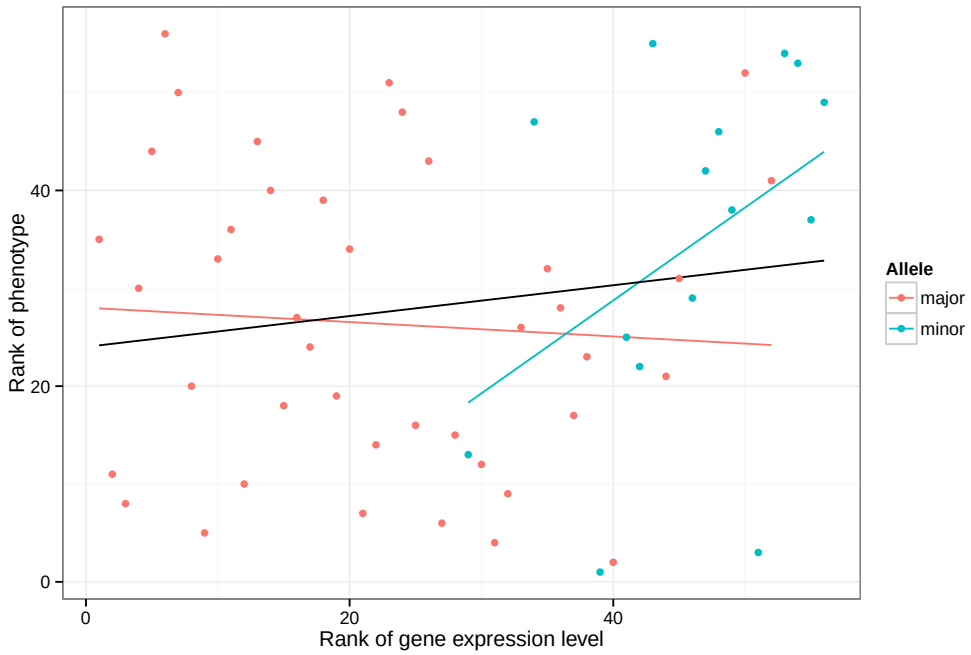


Figure 6.4.: Rank of centroid size of male adult flies plotted against the rank of gene expression level of the gene *Orct2* at 6–8 h post-fertilisation. Points coloured by whether lines were homozygous for the major or minor allele. Coloured lines show linear fits for the respective alleles, black line shows fit through all points. Samples with heterozygous genotype omitted for simplicity.

6.1.3. Location of eQTLs with respect to gene

Analysing the location of eQTLs relative to their associated gene has already proven helpful for quality control of the eQTL sets (see Chapter 5). However, where eQTLs are located with respect to their gene is also interesting biologically, in particular because of the detailed fine mapping that the LD structure in the DGRP allows me to do. Figure 6.5 shows the relative location of eQTLs as in Chapter 5 (Figure 5.1), but this time only considering the fully filtered set. This plot now shows the expected strong enrichment of eQTLs around the promoter region (5′ end) of their associated gene.

Since I tested variants up to a distance of 50 kb for eQTLs, I also found eQTLs further than 1,000 bp away from their gene, which are not shown in this plot. To summarise these, I categorised each lead eQTL by the type of its location:

1. More than 200 bp upstream of the gene
2. Within 200 bp of the gene's 5′ end

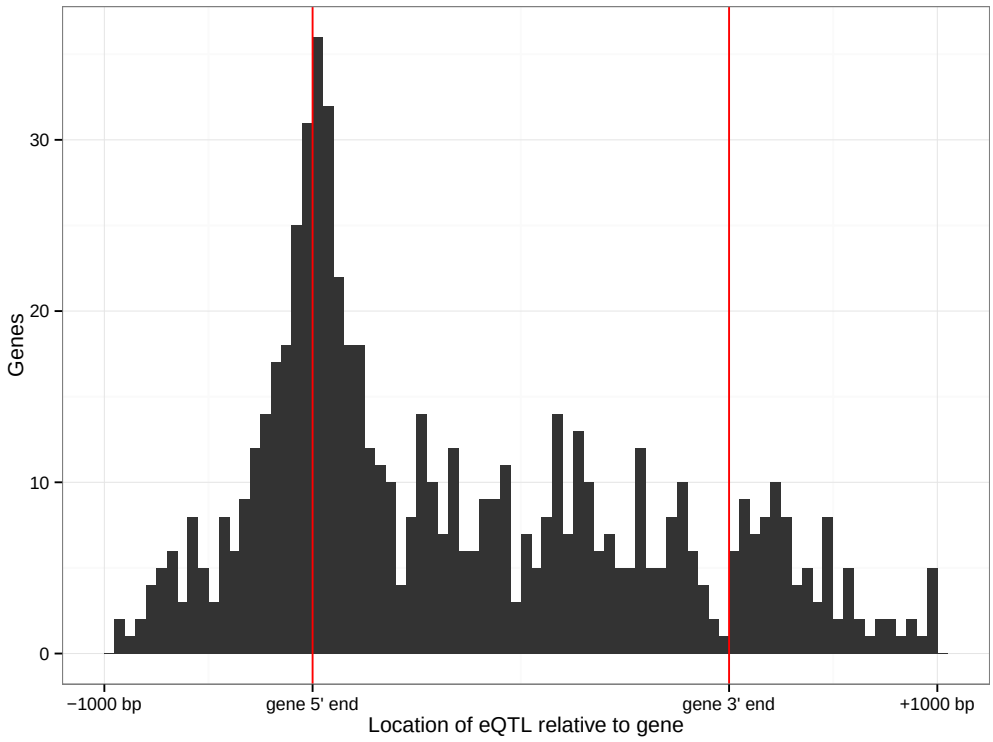


Figure 6.5.: Position of lead gene eQTLs (common effect) relative to their associated gene. Locations inside the gene body in fractions of gene length, locations outside of the gene body in raw base pairs. Red lines denote location of the 5′ and 3′ ends of the gene.

3. In gene body, overlapping an exon
4. In gene body, not overlapping an exon (intronic)
5. Within 200 bp of the gene's 3′ end
6. More than 200 bp downstream of the gene

Figure 6.6 shows the count of lead gene eQTLs with a common effect in each category, compared to a random sample of 100,000 tested variants.

These results show that, while significantly more eQTLs were located close to the 5′ end of genes than would be expected by chance (Fisher's exact test, $p = 1.33 \cdot 10^{-225}$), most eQTLs were actually located further away from the gene, both in the upstream and in the downstream regions. Figure 6.7 shows the absolute distance of the lead gene eQTLs from their associated gene's 5′ end, the putative transcription start site (TSS).

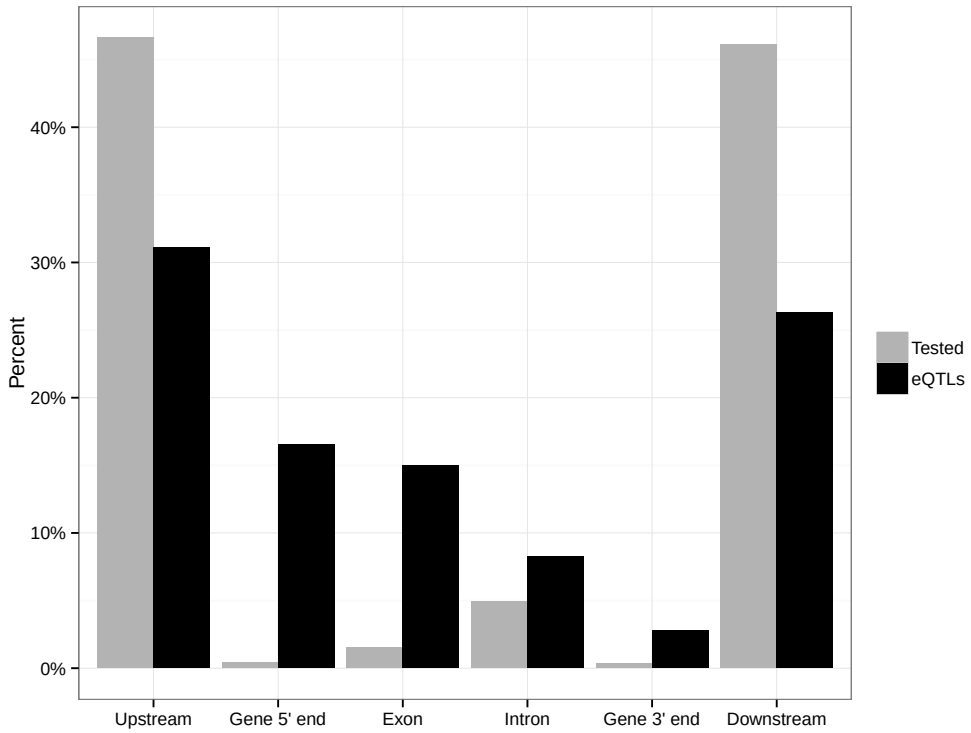


Figure 6.6.: Number of lead gene eQTLs (common effect) in different features of their associated gene, compared to a random sample of 100,000 tested variants. Gene 3′ and 5′ ends include all eQTLs within 200 bp of the respective end.

Again, an enrichment of eQTLs close to the TSS was clearly visible. 415 lead eQTLs (34 %) were located within 1 kb of their gene, 72 % were found within 10 kb. However, I did find eQTLs in the entire gene-proximal 50 kb window I searched, with the furthest being 49,940 bp away. This strongly suggests that there would have been further eQTLs beyond this threshold. I describe my work towards finding these distal eQTLs in Chapter 7.

Genes with their lead eQTL located more than 10 kb away from their TSS were strongly enriched in the GO categories "response to stimulus", "cell proliferation", "single-organism developmental process" and "cell surface receptor signaling pathway" when compared to the genes with closer lead eQTLs (Fisher's test, minimum category size 10, Bonferroni-adjusted p-value < 0.05). This suggests that such genes are more likely to have eQTLs in distal regulatory regions — possibly either because they tend to have more distal regulatory regions in general, or because their promoter-proximal regulatory regions are under stronger evolutionary con-

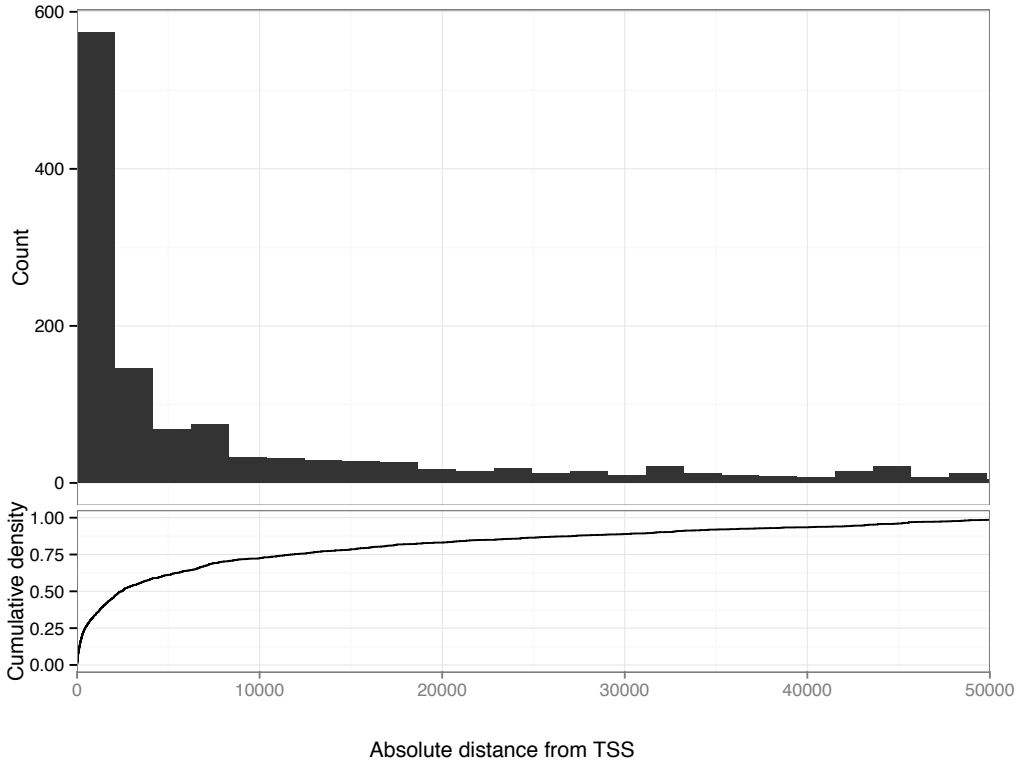
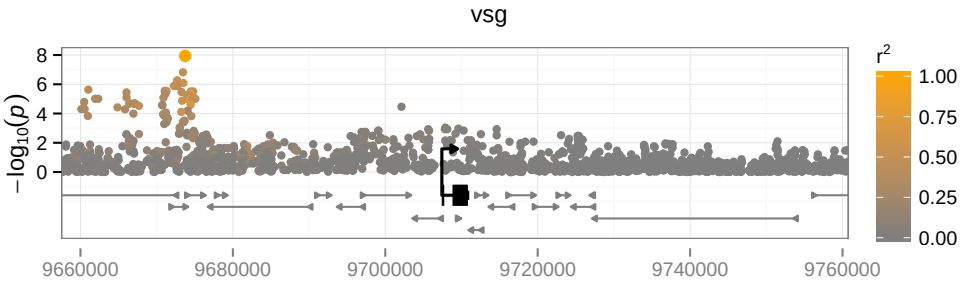


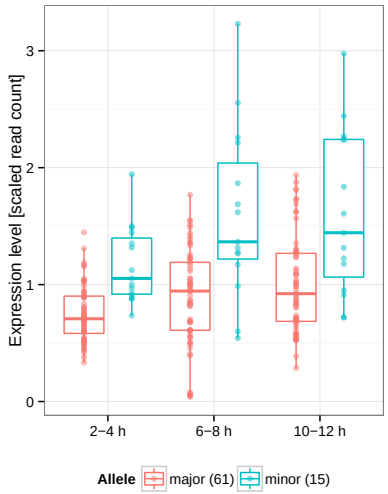
Figure 6.7.: Histogram and cumulative density of the absolute distance of lead gene eQTLs (common effect) from the transcription start site (TSS) of their associated genes. Only the region ±50 kb is shown.

straints. An example of one such gene, *visgun*, is shown in Figure 6.8. The eQTL for this gene is an A to G polymorphism, located more than 37 kb upstream of the TSS.

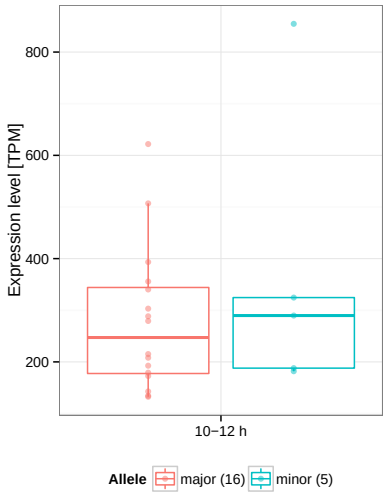
There were seven other genes in the region between this variant and the TSS of *visgun*, six of which I had tested for eQTLs as well. However, none of their expression levels were significantly associated with the genotype at the variant (common effect test, best uncorrected $p = 0.10$), suggesting that this eQTL only has an effect on *visgun*. How this long-range interaction is achieved mechanistically and why the other genes are not affected is not clear, making this a potentially interesting area for further study.



(a) Manhattan plot



(b) 3′ Tag-Seq expression level



(c) RNA-seq expression level

Figure 6.8.: eQTL for the gene *vsg*. (a) Top: Manhattan plot of variants around the gene body (black). Surrounding genes shown as grey arrows. 3′ Tag-Seq coverage not shown. Bottom: Boxplots of 3′ Tag-Seq (b) and RNA-seq (c) expression levels. Samples with heterozygous genotype at the lead eQTL omitted.

6.1.4. eQTLs at the 3′ end of genes

As described in Section 5.2, I applied a filter to all eQTLs that removed any eQTL within 25 bp of a 3′ Tag-Seq peak. This filter did successfully decrease my false positive rate and thus helped me generate a set of high-confidence eQTLs. However, since most 3′ Tag-Seq peaks were located close to the 3′ end of genes, it also led to the removal of almost all eQTLs in that region. It is thus important to note that the low amount of eQTLs at the 3′ end of genes, seen in Figure 6.5, does not tell the whole story. Figure 6.9 shows what Figure 6.5 would look like considering the same mappability filters but omitting the close-to-peak filter.

Without this filter, eQTLs were not only enriched in the 5′ promoter region of

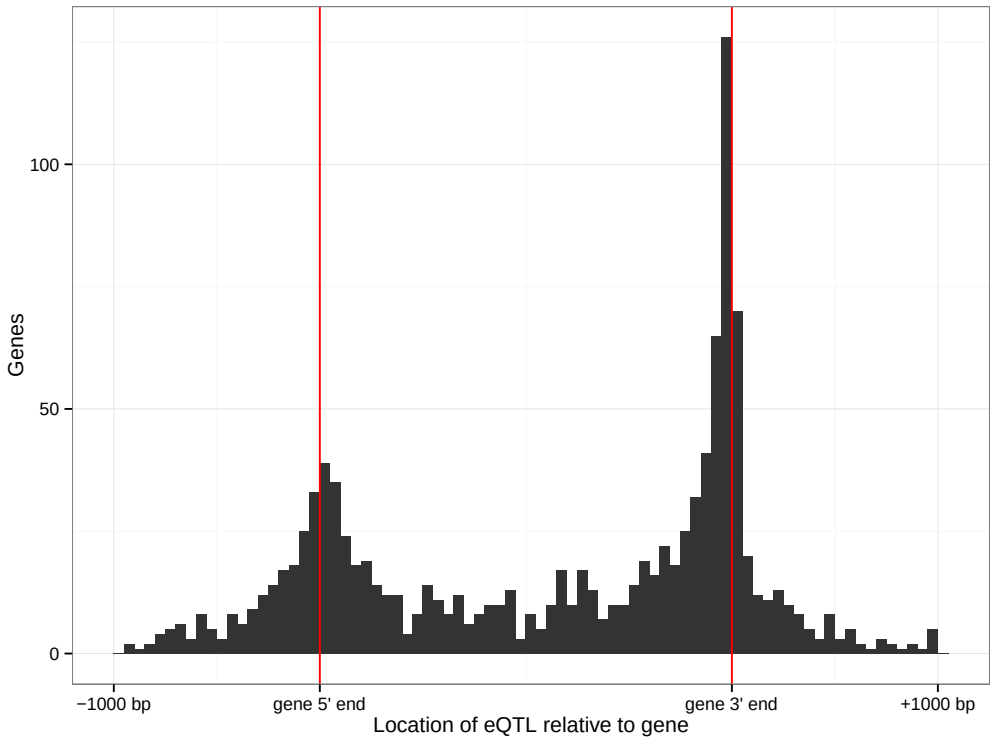


Figure 6.9.: Position of lead gene eQTLs (common effect) relative to their associated gene, not filtered by close-to-peak filter. Locations inside the gene body in fractions of gene length, locations outside of the gene body in raw bp. Red lines denote location of the 5′ and 3′ end of the gene.

genes, but even more strongly enriched around the 3′ end of the gene they were associated with. While some of these eQTLs at the 3′ end were likely to be false positives (see Figure 5.3) some of them were likely to be real eQTLs, affecting processes such as 3′ polyadenylation or post-transcriptional regulation by ncRNAs (see Sections 1.3.2 and 1.3.3) through disruption or creation of sequence motifs.

In particular, I observed a strong enrichment of genes with lead eQTLs located within 100 bp of the 3′ end of 3′ Tag-Seq peaks, the putative site of cleavage and addition of the poly(A) tail. These genes were enriched in GO categories including "cellular component organization" and "sequence-specific DNA binding RNA polymerase II transcription factor activity" compared to all other genes with eQTLs (Fisher's test, minimum category size 10, Bonferroni-adjusted p-value < 0.05). I explore eQTLs that may be associated with polyadenylation in more detail in Section 6.3.

6.1.5. eQTLs in DNase I hypersensitive sites and CRMs

Gene expression and gene regulation are associated with increased DNA accessibility (see Section 1.3.1), which can be estimated using DNase I assays (Gross and Garrard, 1988). In this assay, chromatin is digested by the enzyme DNase I and the digested fragments are sequenced using high-throughput DNA sequencing. As DNase I is much more likely to digest DNA in open chromatin, in particular if that region of the DNA is stressed, the accessibility of a region of DNA can be estimated by the number of sequencing reads obtained from this assay. Regions that are particularly accessible are called DNase I hypersensitive sites (DHSs).

DHSs are known to be markers for regulatory elements such as promoters, enhancers and silencers (Boyle et al., 2008). Thus, eQTLs should be more likely to be located in DHSs than expected by chance. To test whether my eQTLs supported this hypothesis, I overlapped the set of common gene eQTLs with a set of *D. melanogaster* DHSs generated by David Garfield (a post-doctoral fellow from the Furlong group, EMBL-Heidelberg) based on data provided in Thomas et al. (2011). In this test, I only considered the 556 eQTLs that were located further than 1,000 bp from their associated gene, to concentrate my search on putative enhancers and silencers away from the core promoter. Of these 556 eQTLs, 24 % overlapped a DHS, compared to 19 % in a null distribution of randomly chosen variants (one variant per tested gene). This is a significant increase, suggesting that my eQTLs are indeed enriched in DHSs (Fisher's exact test, $p = 0.02$).

In addition, I also overlapped my set of eQTLs with a set of known CRMs from the Redfly database of *Drosophila* regulatory elements (Gallo et al., 2011). As of June 2015, this database contains a collection of 5,557 experimentally validated CRMs. For 1,894 of these, the CRM was also associated with a target gene. While I did not observe a significant enrichment of eQTLs in CRMs in general, I did find 46 genes for which the top eQTL was inside an annotated CRM. For 4 of these the CRM was also assigned to the correct gene, for 36 the CRM was not assigned to any specific gene and for 6 the gene did not match. This mismatch is however not surprising, as CRMs have been shown to have effects on multiple genes at the same time (Link et al., 2013). In fact, these results suggest that my eQTL data set may be useful to annotate these known CRMs further with novel target genes.

6.1.6. Kmer enrichment of eQTLs

The low degree of LD in the DGRP (see Section 1.9) often allowed me to not only determine whether a gene had an eQTL, but also narrow the association down to one or a few likely causal variants. While I would of course still miss ungenotyped variants, the fact that the genotype data that I used had been generated from full genome sequences meant that I was highly likely to have tested the true causal variant, as long as it was biallelic and present in the population with a MAF of at least 5 %.

In particular, the eQTLs where the "eQTL cloud" had a size of one (i.e. where there was no second-best variant within an order of magnitude of the top variant) constituted a set of very likely causal eQTLs. I took the subset of these eQTLs where that single variant was also a SNP to obtain a set of 527 expression quantitative trait nucleotides (eQTNs).

For each eQTN, I extracted the two bases upstream and two bases downstream of the nucleotide from the correct strand of the reference genome, to form kmers of length 5 bp. Using the information whether the eQTN had a positive or a negative effect size I then generated two sets of 527 kmers, one that contained the eQTN with the more highly expressed allele and one that contained the one with the more lowly expressed allele.

I counted how often I observed each kmer in the higher and in the lower set and tested for a significant difference in direction using a binomial test. I opted to use this approach instead of a *de novo* motif enrichment analysis in regions flanking the eQTN (Ettwiller et al., 2007; Bailey et al., 2009; Heinz et al., 2010) for two reasons. First, this approach would not have taken advantage of the fact that I actually know the exact nucleotide that has been mutated, rather than just the general region. Second, motif enrichment tools are usually optimised for ChIP-seq experiments and thus assume that there are few canonical motifs that can explain a large fraction of the signal. This is very unlikely to be the case here, since I did not limit my analysis to a single transcription factor with a single binding site motif.

I observed the strongest difference for the kmer TTCTT which was associated with decreased gene expression levels for five genes but never associated with increased gene expression levels (unadjusted $p = 0.0625$). I further noticed that in the group of the four related kmers TTATT, TTCTT, TTGTT and TTTTT a G or C base ([GC]) in this location was always associated with a decrease in expression (8 decreasing versus 0 increasing cases) while an A or T base ([AT])

was almost always associated with an increase (1 versus 9 cases). The eQTNs for the 8 genes where the change between a TT[CG]TT kmer and a TT[AT]TT kmer resulted in an increase of expression are shown in Table 6.4.

Gene	eQTN	MAF	Location	Low	High
Atf6	2R:1011634	0.13	upstream	C	T *
CG15369	X:9094032	0.11	upstream	C *	T
CG15544	3R:26637326	0.06	upstream	G *	T
PI31	2R:7881876	0.08	upstream	C *	T
Prp38	2R:5018473	0.05	upstream	C	T *
CG13133	2L:10061227	0.08	downstream	G *	A
CG42342	3R:12350405	0.27	internal intron	C *	T
Cpsf73	3R:14489189	0.16	internal exon	C	T *

Table 6.4.: eQTNs disrupting the TT[CG]TT motif with a corresponding increase in expression. Low, nucleotide in the more lowly expressed allele. High, nucleotide in the more highly expressed allele. Major allele marked with an asterisk (*).

For 5 out of 8 genes, the variant was located a considerable distance (10–25 kb) upstream of the TSS. For the sixth case, CG13133, the eQTN was 30 kb downstream of the 3′ end. The eQTN for the seventh gene, CG42342, was located in the middle of the annotated gene region. However, this gene is almost 70 kb long and has many exons, suggesting that the kmer may actually still have been located several kb upstream of an unannotated promoter for a shorter isoform. Finally, for *Cpsf73*, the kmer was located directly in the promoter region, 61 bp inside the gene body. The fact that at least six of the kmers were located so far away from the TSS, with a strong bias to the upstream region, suggests that these kmers may be an important part of a distal CRM.

Intriguingly, TTGTT or its reverse complement AACAA is known to be part of the consensus binding site sequence of transcription factors containing a SOX domain, including *Dichaete* in *Drosophila* (Ma et al., 1998; Noyes et al., 2008) and *Sox5* in mouse (Denny et al., 1992). While it is important to stress that this analysis is based on very few observations, these eQTNs may thus prove an interesting target for further study.

6.1.7. Negative selection and the Winner's Curse

In theory, I would expect that eQTLs leading to a large change in gene expression would be selected against, since a big difference in the expression level of a gene is likely to be detrimental. Thus, strong eQTLs should generally be under negative

selection, leading to a low frequency of the allele in the population. In order to determine whether I could see such an effect in my data, I plotted the absolute effect size of each lead gene eQTL (common effect) against the MAF of the associated variant (Figure 6.10). If negative selection was acting on the eQTLs, I expected to see a depletion of strong effect sizes at high MAF and an enrichment of strong effect sizes at low MAF.

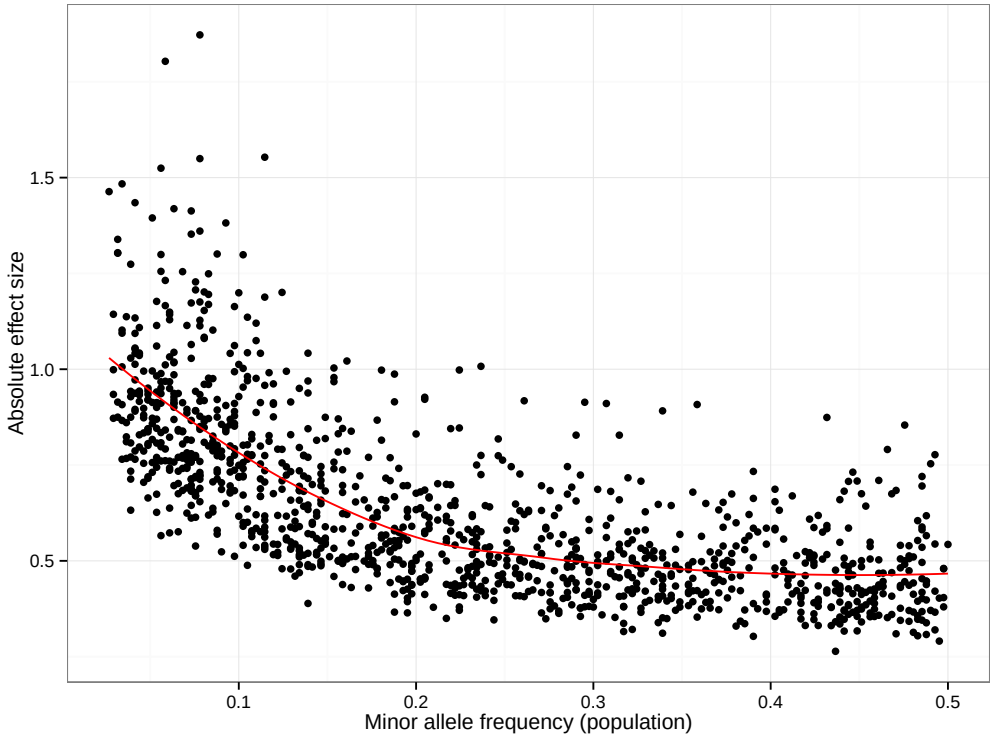


Figure 6.10.: Absolute effect size of lead gene eQTLs (common effect) plotted against the MAF of their associated variant. Red line shows a LOESS fit.

Indeed, this plot seemed to show evidence of negative selection, with most strong eQTLs having a low MAF, and no high-frequency variant exhibiting strong effects. If the eQTLs were not under selection, I would have expected the effect sizes to be randomly distributed across the allele frequencies.

However, it soon emerged that this effect was almost certainly a manifestation of the Winner's Curse (Button et al., 2013; Halsey et al., 2015). In short, this describes the phenomenon that studies identifying new effects almost always overestimate the size of that effect, in particular when they are lowly powered. The power to detect eQTLs is correlated with the MAF of the tested variant as I have

shown when I calculated the statistical power of this study (Section 4.3). Thus, it was possible that, due to the Winner's Curse, I not only overestimated effect sizes for some eQTLs but was also more likely to do so for variants with a low MAF.

To illustrate this problem, I simulated this effect for an example eQTL. I generated data reflecting a change in expression level with a constant effect size of 0.25, to which I added noise from a standard normal distribution ($\mu = 0$, $\sigma = 1$) to reflect experimental and biological variation. I generated 100,000 variants with exponentially decreasing MAF and tested each of them for association with the effect. Figure 6.11 shows the effect size estimates for each of the 100,000 variants. For high MAF, the estimated effect size was close to the true value of 0.25 (blue dashed line), with little variation around it. However, for lower MAF, the spread around the real effect size was much larger, with estimates reaching values of more than ± 3.

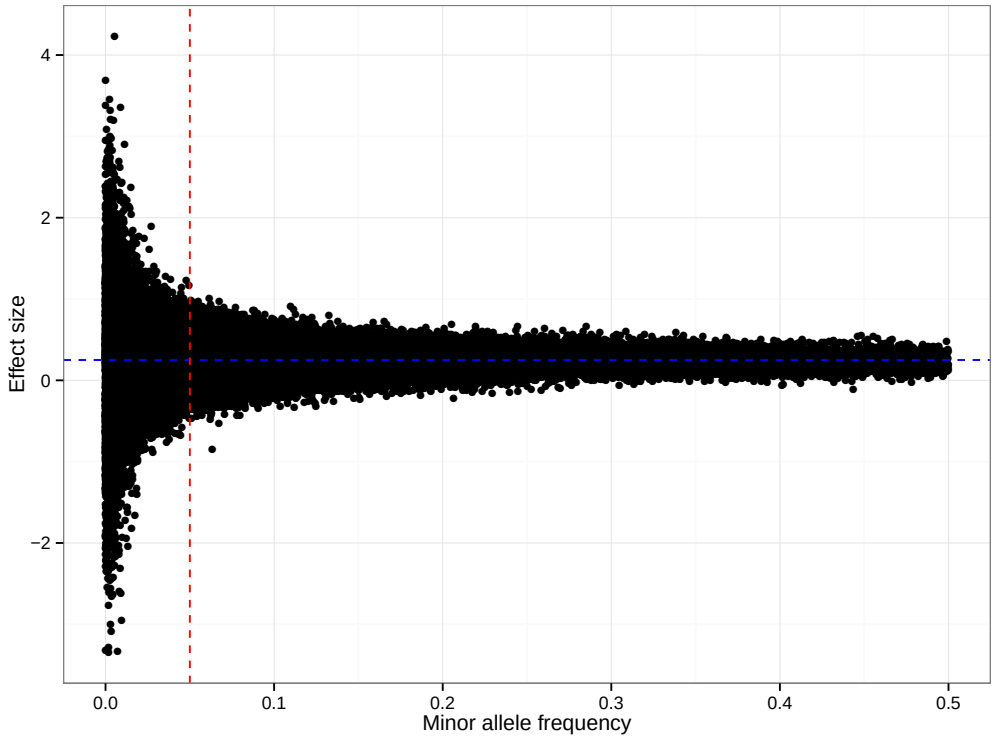


Figure 6.11.: Estimated effect size of simulated eQTLs plotted against the MAF of the variant. Blue dashed line shows the true effect size of $\beta = 0.25$, red dashed line shows the MAF cutoff of 5 %.

I adjusted the p-values for multiple testing using BH's method and extracted the 2,399 variants with MAF $> 5\,\%$ that passed a strict threshold of $1\,\%$ FDR. The resulting distribution of absolute effect sizes is shown in Figure 6.12. Only the variants for which the effect size happened to be larger than the real effect passed the p-value threshold. The lower the MAF was, the more pronounced this overestimation of the effect size became, due to the larger error of the estimates.

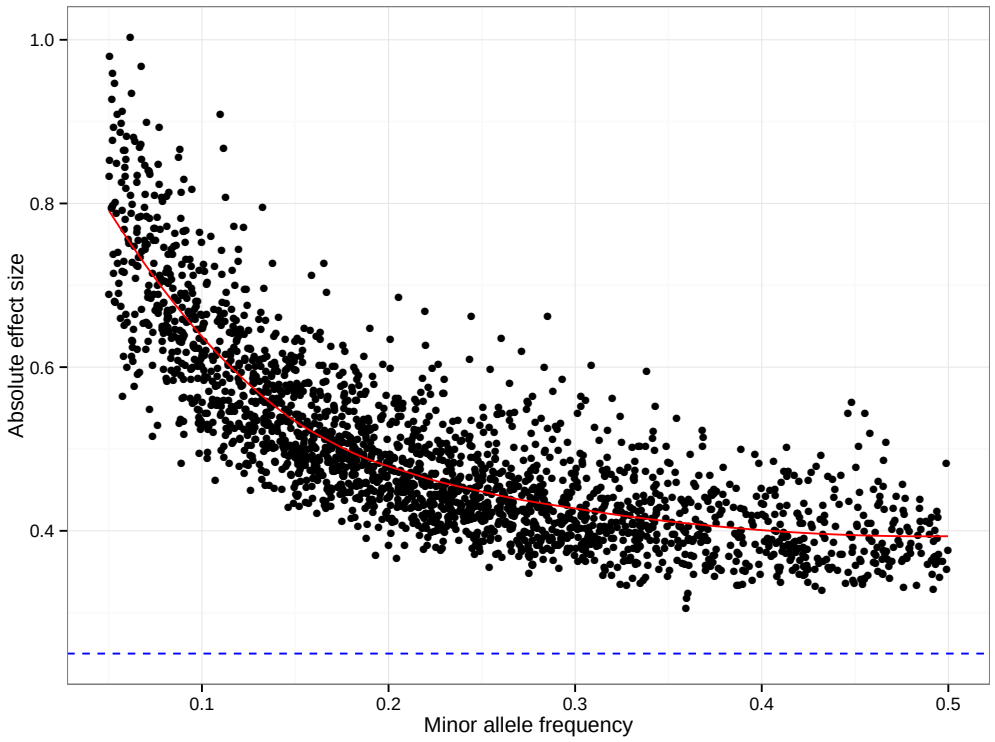


Figure 6.12.: Absolute effect size of simulated eQTLs plotted against the MAF of the variant. Only variants with FDR $< 1\%$ and MAF $> 5\%$ are shown. Blue dashed line shows the true effect size of $\beta = 0.25$, red line shows a LOESS fit.

The clear resemblance between the simulated (Figure 6.12) and the real data (Figure 6.10) suggests that this supposed negative selection effect is indeed likely to be a result of the Winner's Curse. It is still possible that my observations may have been caused by a combination of this artefact and true biological effects, but I was unable to disentangle these effects based on this data.

6.2. Common and stage-specific gene eQTLs

One of the main novel aspects of my study is the measurement of gene expression levels at three different developmental stages rather than just one. The expectation was that this would allow me to find not only eQTLs that had an effect whenever the gene was expressed, but also eQTLs that only had an effect during specific stages of development. Mechanistically, these eQTLs might, for example, be located in CRMs that are bound by a transcriptional regulator that is only active during a single stage in development.

Based on the classification procedure described in Section 4.6.2, I found common eQTLs for 1,224 genes, eQTLs specific to 2–4 h for 136 genes, eQTLs specific to 6–8 h for 79 genes, eQTLs specific to 10–12 h for 113 genes, 20 complex cases and 177 weak two-stage eQTLs. A heatmap of the absolute effect sizes of specific eQTLs compared between the three developmental stages is shown in Figure 6.13. The effect sizes shown here were obtained from the single-stage eQTL testing, as described in Section 4.6.2.

This plot illustrates the complex interaction between development and gene regulation, with all possible combinations of stage-specific effects being observed.

6.2.1. Assigning each eQTL to a developmental stage

In addition to looking for eQTLs that were specific to a single developmental stage, I also classified each eQTL by the stage in which the effect was most pronounced. For this classification, I fitted an additional multivariate model in LIMIX, which allowed for a different effect size coefficient β_i to be assigned to each developmental stage:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \otimes \mathbf{1}_N + \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \otimes \mathbf{x} + \mathbf{g} + \boldsymbol{\psi} \tag{6.1}$$

With $\mathbf{y}_i, \mu_i, \mathbf{1}_N, \mathbf{x}, \mathbf{g}$ and $\boldsymbol{\psi}$ as defined in Section 4.5. For each variant that was found to be significantly associated with the common or specific effect tests, I extracted the values β_1, β_2 and β_3 from this model and assigned the eQTL to the developmental stage i where β_i was largest. I found 661 genes with their strongest eQTL at 2–4 h, 515 genes with their strongest eQTL at 6–8 h and 573 genes with their strongest eQTL at 10–12 h after fertilisation.

Using this annotation, I compared the relative location of each eQTL to the

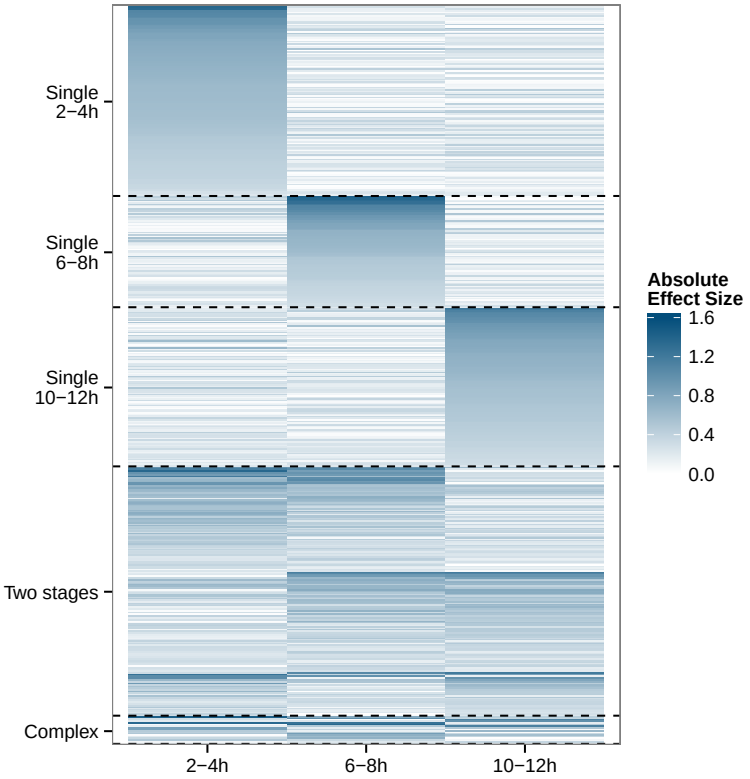


Figure 6.13.: Single-stage effect sizes of the 525 lead eQTLs with different specificity (y-axis) in the three developmental stages (x-axis). Sorted by absolute effect size within each group.

associated gene between the different developmental stages. As shown in Figure 6.14, the eQTL locations were largely similar between developmental stages.

Interestingly, there did seem to be an increase in the amount of eQTLs inside of the gene body at 2–4 h. In particular exonic eQTLs were almost twice as common at 2–4 h as at 6–8 h and 10–12 h (122 versus 47 + 60 cases, Fisher's exact test, $p = 3.92 \cdot 10^{-7}$). Among the genes with an exonic eQTL strongest at 2–4 h was the developmental gene *ecd*, which I had identified earlier.

One possible way mutations in the exon of a gene can affect its expression level is through changes to miRNA binding sites (see Section 1.3.3). 2–4 h is the developmental time point during which the maternal-to-zygotic transition (MZT) occurs, which has previously been associated with regulation by miRNAs in *Drosophila* as well as other organisms (see Sections 1.2.1 and 1.3.3).

To gain a better understanding of this phenomenon, I performed a second kmer

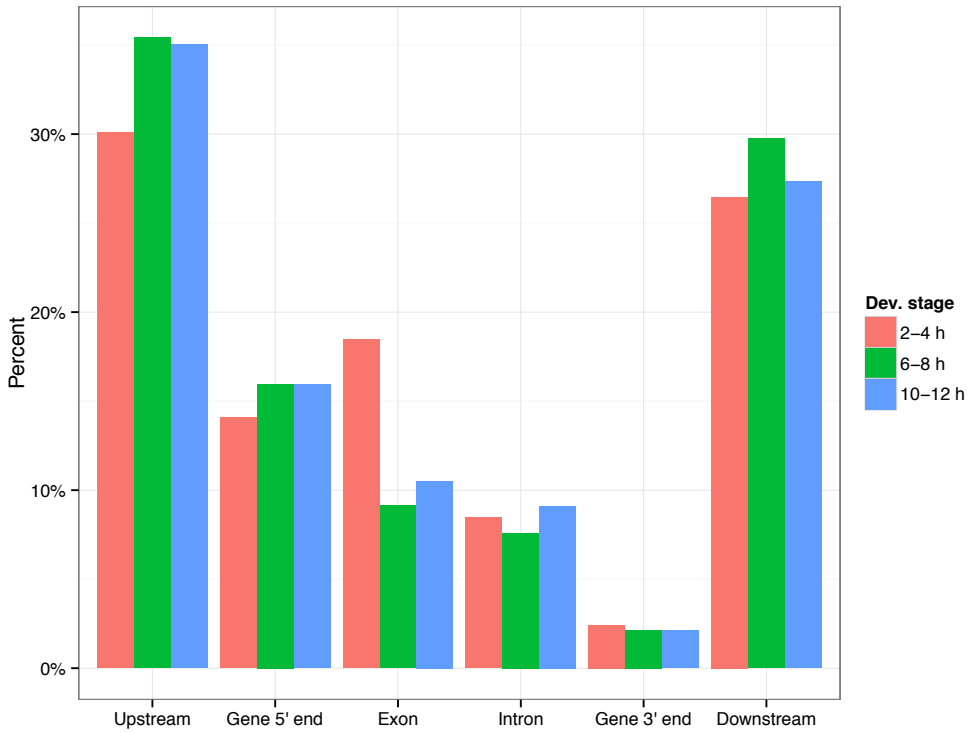


Figure 6.14.: Percentage of lead gene eQTLs (both common and specific effect) in different features, coloured by the developmental stage (Dev. stage) in which the strongest effect was observed. Gene 3′ and 5′ ends include all eQTLs within 200 bp of the respective end.

enrichment analysis (as described in Section 6.1.6) on the 55 exonic eQTNs that were associated with a single SNP and had their strongest effect at 2–4 h. Searching for a kmer of length 5 bp within 2 bp of the eQTN I found that the kmers GC[AT]GG appeared to be associated with a decrease in gene expression, with 10 negative and 2 positive associations. I performed the same analysis on the 61 exonic eQTNs that had their strongest effect in 6–8 h or 10–12 h and did not find a single instance of this kmer in these stages.

This hints at the presence of a gene regulatory mechanism associated with the kmer GC[AT]GG that is more active in early stages of development than in later ones. A possible explanation for this would be that there is a miRNA that targets this sequence, which is only expressed at 2–4 h, potentially because it is involved in the MZT. Exonic mutations have also been observed to affect splicing (Cartegni et al., 2002), and sequences associated with this phenomenon in humans have

indeed been observed to contain the reverse complement sequences of the kmer, CCTGC and CCAGC (Cavaloc et al., 1999). However, due to the low number of cases I had observed, I was unable to investigate this further.

6.3. 3′ isoform eQTLs and alternative polyadenylation QTLs

So far, I have only described eQTLs that affected the total expression level of genes. However, as I described in Chapter 4, I also tested each 3′ Tag-Seq peak individually for association with proximal variants. This allowed me to find a set of eQTLs that affected the usage of individual 3′ isoforms of a given gene, which I called 3′i-eQTLs.

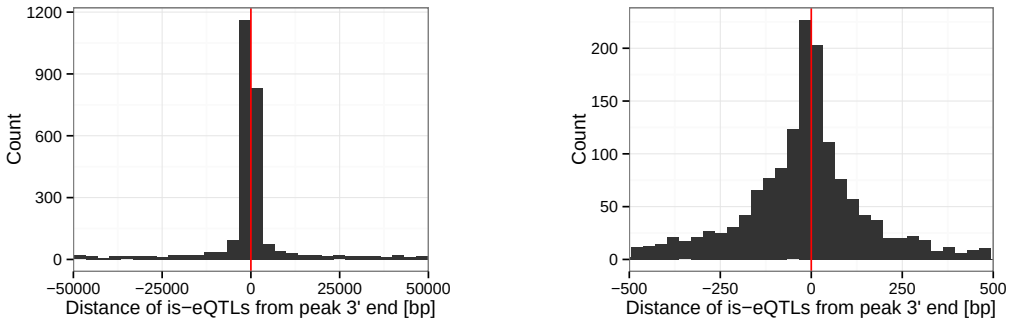
Several mechanisms have been suggested that may result in alternative polyadenylation (APA) through changes to different regions association with the polyadenylation procedure (see Section 1.3.2). This includes the canonical poly(A) motif itself (AAUAAA) as well as the USE and DSE. Changes to the motifs located in these regions may increase or decrease the rate of cleavage, polyadenylation and subsequent termination at a 3′ cleavage site. If these processes are changed at the first poly(A) site, this may then have the opposite effect on downstream 3′ ends as transcription continues to the next poly(A) site. Depending on how this change in poly(A) site usage affects the mRNA stability and degradation rate, this change may or may not lead to a change in the overall gene expression level, reflected in the steady-state level of mRNA.

As I described in Chapter 5, I applied the same set of filters to the 3′i-eQTLs as for the gene eQTLs, except for the close-to-peak filter which would have removed almost all of the data. This difference results in one important caveat with the analysis of 3′i-eQTLs: the potential for mappability artefacts as described in Section 2.2 is larger than for gene eQTLs. I applied the same stage-specificity classification as described in Section 6.2 to the 3′i-eQTLs, resulting in a set of 5,883 common 3′i-eQTLs and 1,241 stage-specific 3′i-eQTLs. However, in the following sections, I will only discuss the 5,883 common 3′i-eQTLs.

For genes with only a single 3′ Tag-Seq peak, the 3′i-eQTLs were almost equivalent to the associated gene eQTLs. This was expected, as the gene expression level is calculated as the sum of all 3′ Tag-Seq peak expression levels. However, for genes with more than one observed 3′ end, the 3′i-eQTLs could exhibit more interesting behaviours.

6.3.1. Isoform-specific eQTLs

Particularly interesting were the cases where a variant was associated with the expression level of an individual 3′ isoform, but not of the whole gene. Such eQTLs could be specifically affecting the usage of a single poly(A) site and might thus allow me to observe the effects of a variant on 3′ polyadenylation. As a first step, I thus removed all 3′i-eQTLs where the associated gene also had a common eQTL, to obtain sets of 3′i-eQTLs where the eQTL was not associated with the overall gene expression level. Of the 5,883 peaks with at least one common 3′i-eQTL, 2,682 did not have any significant common effect eQTL for their associated gene. I called these 2,682 3′i-eQTLs, associated with 1,689 different genes, isoform-specific eQTLs (is-eQTLs). Many of these is-eQTLs were located closely upstream or downstream of the 3′ end of their associated peak, with a median distance of 22.5 bp upstream (Figure 6.15).



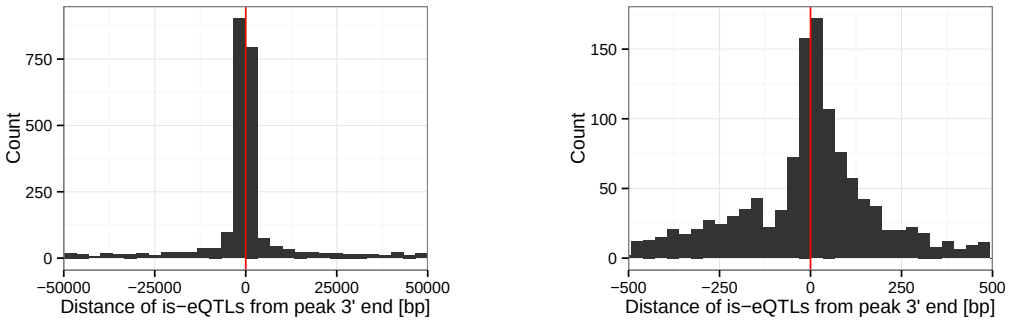
(a) Full **(b)** ±500 bp

Figure 6.15.: Distance of is-eQTLs to the 3′ end of the associated 3′ Tag-Seq peak. (a) Full histogram. (b) Histogram of all is-eQTLs within 500 bp of the 3′ end.

As the 3′ end of the 3′ Tag-Seq peak is approximately equivalent to the cleavage site, this would be consistent with the hypothesis that many of these is-eQTLs are affecting 3′ APA signals, such as the canonical poly(A) motif AAUAAA or the USE and DSE upstream and downstream of it. However, this location could also be a sign of artefacts caused by mappability problems, with variants inside the peak region affecting how many reads could be mapped back to the peak in different individuals. While I did attempt to account for these problems (see Sections 2.2 and 5.2) it was still difficult to confidently state whether or not these is-eQTLs were the consequence of real gene regulatory mechanisms.

In an attempt to shed more light on this, I determined the location of the

summit of each peak, calculated as the median among the summit locations in each individual sample. Since the read length we used for 3′ Tag-Seq was 43 bp and I only considered the height of the summit to estimate the expression level of the peak, any variant further than 43 bp away from the summit should not result in changes to the estimated expression level. After removing all is-eQTLs within 43 bp (the read length) of the peak summit there was still a visible enrichment of is-eQTLs around the 3′ end of the peak. The median distance of is-eQTLs to their associated 3′ end was now 11 bp upstream, as shown in Figure 6.16.



(a) Full **(b)** ±500 bp

Figure 6.16.: Distance of is-eQTLs further than 43 bp away from their 3′ Tag-Seq peak's summit to its 3′ end. (a) Full histogram. (b) Histogram of all is-eQTLs within 500 bp of the 3′ end.

In theory, we could imagine that each is-eQTL switches the 3′ end usage of a given transcript between two alternative 3′ isoforms. In one allele, the poly(A) machinery recognises the first set of poly(A) motifs and begins the process of polyadenylation, resulting in the first 3′ isoform being used. In the other allele, a variant decreases the efficiency of polyadenylation of the first motif and transcription continues to the next poly(A) site, where a new 3′ end is generated. The result would be that the usage of the first poly(A) site is decreased or entirely abolished, with a corresponding increase in usage of the second poly(A) site.

Thus, every peak with an is-eQTL around its poly(A) motif should be followed by another 3′ end peak. I tested this assumption for the set of 620 is-eQTLs that were located within 100 bp of the 3′ end of their associated peak, but at least 43 bp from its summit, suggesting that they affected cleavage at the given 3′ end. For each of these is-eQTLs, I determined whether there was a second 3′ Tag-Seq peak associated with the same gene located further downstream. 512 (83 %) is-eQTLs had a second 3′ Tag-Seq peak further downstream. Overall, only

18,993 (71 %) of tested 3′ Tag-Seq peaks associated with multi-peak genes had a second peak further downstream, which is significantly fewer (Fisher's exact test, $p = 1.28 \cdot 10^{-10}$). These results are consistent with the assumption that these is-eQTLs are affecting the efficiency of cleavage and polyadenylation of the 3′ end, resulting in a shift in the 3′ end usage to other sites further downstream.

As in Section 6.1.6, I searched for 5-mers affected by is-eQTLs with an excess of positive or negative effects on the expression level. I tested the 453 is-eQTLs that were located within 100 bp of the 3′ end of their peak, at least 43 bp away from their peak's summit and with another peak further downstream. From these, I again considered only the 319 is-eQTLs that were associated with only a single SNP, which meant that I was likely to have identified the causal is-eQTN.

The most strongly enriched kmers were ATAAA and AAAGA, which were both observed nine times on the more highly expressed allele and never on the more lowly expressed allele. In both cases, an A-nucleotide in the more highly expressed allele had been replaced by another nucleotide in the more lowly expressed allele (AT[CGT]AA and AA[CGT]GA). Both of these kmers are similar to the canonical poly(A) motif AATAAA, with the first kmer actually being a substring of it. Thus, these may be cases where changes to the canonical poly(A) motif have led to APA, with the binding affinity of the poly(A) machinery differing between the two alleles. Overall, is-eQTLs seemed to be heavily enriched in A-rich regions.

6.3.2. Alternative polyadenylation QTLs

The previous analysis had already shown that peaks with is-eQTLs at their 3′ end tended to be followed by another peak, in line with their expected involvement in APA. In fact, each such pair of peaks should result in two is-eQTLs, as any variant associated with a decrease in the usage of the first peak should also be associated with an increase in the usage of the second peak. An example of such 3′ isoform switching, which I observed for the gene *YL-1*, is shown in Figure 6.17.

The difference in transcript length between the major and minor allele due to this is-eQTL is clearly visible, both in the 3′ Tag-Seq and in the RNA-seq data. The SNP associated with this switch is located inside the first 3′ Tag-Seq peak, 22 bp upstream of the putative cleavage site, in the region expected to contain the canonical poly(A) motif (see Section 1.3.2). While individuals with the major allele have the sequence GAATAAAA at this location, individuals with the minor allele have the sequence GAATAATA instead. Thus, the is-eQTL disrupts the canonical AATAAA motif at the first poly(A) site, decreasing its usage. The

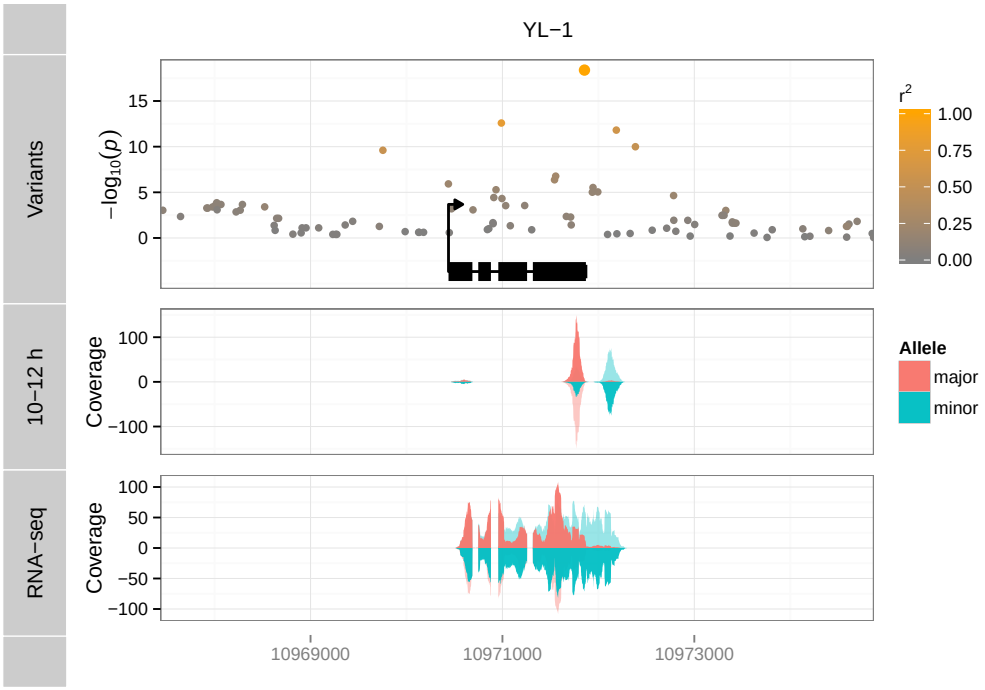


Figure 6.17.: 3′ isoform switching of the gene *YL-1*. Top: Manhattan plot of variants around the gene body (black). Middle: Median 3′ Tag-Seq coverage at 2–4 h, 6–8 h and 10–12 h, grouped by major/minor allele at the lead eQTL. Only reads associated with *YL-1* shown. Bottom: Median RNA-seq coverage at 10–12 h, grouped by major/minor allele at the lead eQTL. Only reads associated with *YL-1* shown.

location of the 3′ end consequently shifts to the second poly(A) site, approximate 350 bp downstream.

In practice, several factors made the comprehensive identification of these cases more complicated. First, the rate of mRNA degradation can differ between 3′ isoforms, for example through the inclusion of different miRNA binding sites (see Section 1.3.3). Consequently, the steady-state expression levels may have differed if alternative 3′ polyadenylation sites were used, even if the transcription rate itself stayed the same. Second, if a gene had more than two 3′ isoforms, the signal became challenging to detect: When one of these 3′ isoforms was decreased by an is-eQTL, the corresponding increase could become split up among the other isoforms, resulting only in small changes to their expression levels. Finally, any random variation in the testing or problems with the mappability of one of the peak regions may have led to one of the is-eQTLs not being called. Hence, while I did find at least two is-eQTLs with opposite signs for 356 out of 1,689 genes

with is-eQTLs, only 43 of them involved the same lead variant. I thus searched for cases of 3′ isoform switching in a more systematic way, without relying exclusively on the is-eQTLs.

For this purpose, I returned to the full set of significant common 3′i-eQTLs and identified those cases where the associated gene had at least two 3′ Tag-Seq peaks. For each time point, I determined the most strongly expressed 3′ isoform of each gene in each line based on the raw 3′ Tag-Seq expression level. Counting how often each peak was the most highly expressed peak for each of the genotypes, I then built a $N \times G$ contingency table for each gene, where N is the number of peaks and G is the number of genotypes. I tested for differences in 3′ peak usage patterns between the genotypes using a two-sided Fisher's exact test and calculated the FDR (across tested variants, developmental stages and genes) using the BH method.

I applied a stringent threshold of FDR $< 1\,\%$ and removed all cases where one of the peaks would have failed the mappability filters described in Section 5.2. In addition, I removed all cases in which the most commonly used 3′ isoform was the same in both the major in the minor allele. This resulted in a final set of 405 loci associated with alternative polyadenylation of 186 genes, which I called apaQTLs. For each gene, I selected the apaQTL with the best p-value for a final set of 186 lead apaQTLs. A plot of the distances between the most used 3′ ends at the major and the minor allele for 127 apaQTLs is shown in Figure 6.18. There were 59 genes with an apaQTL that led to a change of more than 1,000 bp, which I have omitted from this plot.

Interestingly, for 73 genes with an apaQTL, the variant was only significantly associated with a change in poly(A) site, but not with a change in expression level. This suggests that the steady-state expression level for this gene stayed relatively constant, regardless of the poly(A) site that was being used. While these different 3′ isoforms thus did not appear to be differentially degraded, they may have had different translation rates because of differences in their 3′ UTR or yielded different proteins through alternative splicing coupled with APA resulting in the omission or addition of exons.

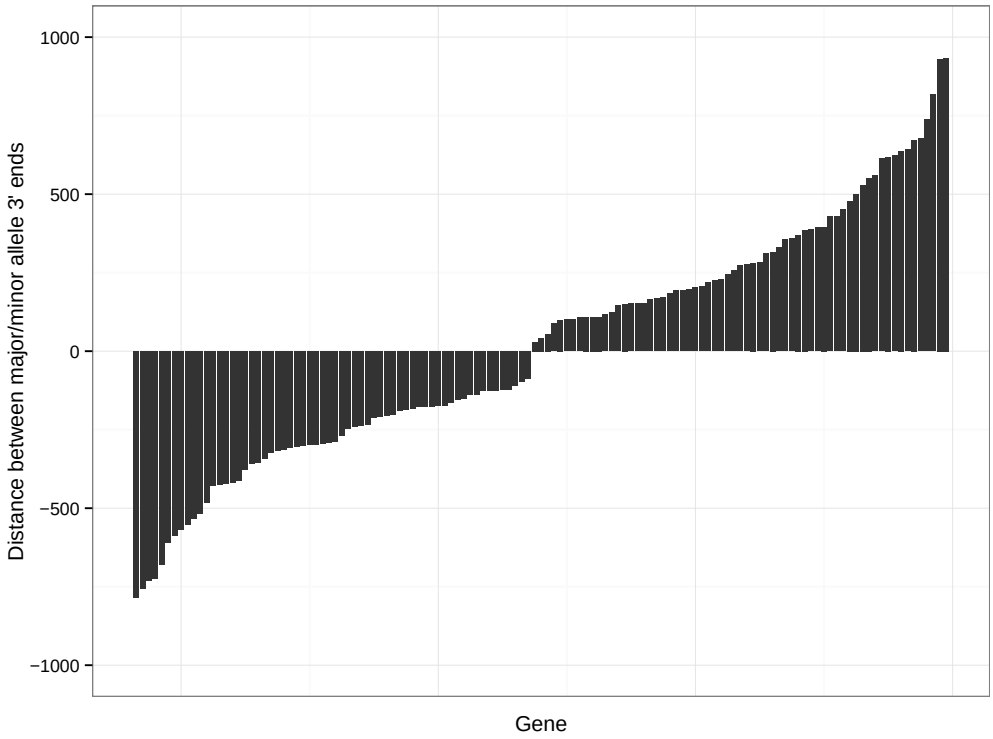


Figure 6.18.: Distance between the most used 3′ ends at the major and the minor allele for 127 apaQTLs. Genes with a distance of more than 1,000 bp omitted.

7. Distal and *trans* eQTLs

In Chapters 4 and 5 I described how I searched for gene-proximal eQTLs, located within 50 kb of the gene or 3′ transcript end that they were associated with. In this chapter I describe how I extended this approach to search for eQTLs genome-wide, and investigate some properties of distal and *trans* eQTLs.

7.1. Introduction

I had based my choice of window size for the gene-proximal eQTL search on the observation that over 99 % of annotated *D. melanogaster cis*-regulatory modules (CRMs) are located closer than 50 kb to their gene, as described in Chapter 4. By focusing on this region close to the gene, I concentrated on variants that were most likely to be eQTLs, while ignoring the vast majority of other variants that were unlikely to be associated. This decreased the computational complexity of the testing and increased the power to detect effects, since I only needed to test an average of 1,392 variants per gene instead of all 1,772,891 variants genome-wide.

The analysis of the proximal eQTLs (see Chapter 6) suggested that this was a valid assumption, with more than 80 % of the proximal eQTLs located within 15 kb of the gene. However, I also observed eQTLs further away from their associated gene, throughout the entire search window of up to 50 kb.

This showed that, while the vast majority of eQTLs could indeed be found proximal to the gene, I was potentially missing some interesting eQTLs further away. These distal eQTLs may represent long-range *cis* effects as well as *trans* effects. In addition, I was also interested in finding eQTLs on different chromosomes, which would be likely to be true *trans* effects.

In this chapter, I describe the process and the results of mapping eQTLs genome-wide, with the aim of finding distal *cis* and *trans* eQTLs. For the purpose of these analyses, I consider an eQTL to be in *trans* if it is located on a different chromosome arm from its associated gene.

7.2. Genome-wide calling of eQTLs

As for the gene-proximal eQTLs (Chapter 4), I tested for common effects using a multi-stage linear mixed model and processed the results into a set of genome-wide common eQTLs. I did not attempt to call stage-specific eQTLs as I did not expect the statistical power to be sufficient to find such effects.

Conducting such a GWAS on 10,094 phenotypes (the expression levels of each gene) posed some major computational challenges.

For proximal eQTLs, I had calculated empirical p-values to account for multiple testing and local LD structure of each gene, using 10,000 permutation experiments per gene (see Section 4.5). It was possible to calculate this in less than an hour per gene, as each permutation took less than a second to calculate. However, for the genome-wide eQTLs, this number would have increased to approximately 10 CPU minutes per permutation and 60 CPU days per gene, for a total of approximately 1,658 CPU years.

Thus, for the genome-wide eQTL testing, I opted to control the FDR across all tested variants and genes at the same time using the BH method instead. This had the advantage that I did not need to perform any permutation experiments, vastly increasing the speed of my computations but at the cost of decreased statistical power. As early trials indicated that the BH method was much more sensitive to outliers caused by non-normally distributed phenotypes than my previous approach, I adjusted the gene expression level normalisation procedure described in Section 3.5 as follows to make the data more suitable for genome-wide eQTL calling.

Instead of simply centring and scaling the gene expression levels within each gene, I quantile-normalised them by transforming them into a standard normal distribution (inverse normal transformation). I used PEER (see Section 3.5.1) on these normally distributed expression levels to capture residual variation while correcting for up to $k = 25$ hidden confounding factors (Section 7.2.1 describes how this number was chosen). Finally, I performed a second inverse normal transformation on these residuals to obtain the final, normally distributed phenotypes. This procedure likely further decreased the statistical power by omitting some of the true structure in the data, but also removed outliers from the data which would have otherwise increased the false positive rate.

However, even the BH method could not be applied directly to this data, due to the amount of tests I had performed. I had tested $V = 1{,}772{,}891$ variants for asso-

ciation with $N = 10{,}094$ genes, which resulted in a total of $V * N = 17{,}895{,}561{,}754$ (17.9 billion) p-values to be corrected. Simply holding these numbers in memory would already have required at least 143 gigabytes of RAM (64 bits per p-value). To handle this amount of data in an efficient way, I developed a custom approach that allowed me to apply the BH method to my data set by parallelising parts of the computation and avoiding unnecessary calculations.

This approach took advantage of the fact that BH is a step-up procedure, meaning that it involves analysing p-values in increasing order (starting from the most significant) until a suitable threshold is reached. After this threshold has been found, all p-values greater than it are guaranteed to not pass the FDR threshold. Thus, I was able to decrease the number of p-values I had to test vastly by only considering p-values likely to be significant. I chose $p = 10^{-5}$ as a conservative threshold and extracted eQTLs with p-values smaller than this in a parallelised fashion.

This resulted in a much more tractable set of 415,775 p-values. I sorted these p-values in order of increasing value and calculated the critical p-value for each of them using the formula $c_i = \text{FDR} \cdot \frac{i}{n}$ with FDR = 10%. However, instead of setting n to the length of the p-value vector as in the normal BH procedure, I set it to the original number of p-values $V * N$. I then found the largest $i = i_{\max}$ such that $p_i \leq c_i$ and classified all p-values with rank $i \leq i_{\max}$ as passing the FDR threshold. This procedure is mathematically equivalent to applying the BH procedure to the entire vector of p-values but could be calculated within seconds using a standard amount of RAM.

7.2.1. Optimising the number of hidden factors

A fundamental problem in searching for *trans* effects while also accounting for experimental batch effects is that it is not necessarily possible to distinguish between the two. When I corrected for batch effects using PEER, I assumed that eQTLs should only affect few genes, meaning that any systematic difference between samples that affected many genes at once was likely to be a batch effect. This assumption is not necessarily valid for *trans* eQTLs, since a change in the protein structure or the expression level of, for example, a transcription factor may affect the expression of many downstream genes.

The number of different effects that PEER tries to remove from the data is given by the number of hidden factors k. The larger k is, the more structure may be removed from the data. Experimental noise, which I expect to be stronger than

the effects of *trans*-acting factors, will be removed first. Consequently, k needs to be large enough to capture most of the possible experimental batch effects, but small enough such that *trans* effects are not yet removed from the signal.

I normalised the data as described above, using different values of k between 1 and 50. For each k, I called eQTLs genome-wide and applied the FDR correction using the BH procedure. I then counted how many unique genes had at least one eQTL, as well as how many genes had at least one *trans* eQTL. The result is shown in Table 7.1.

k	eQTL genes, total	eQTL genes, *trans*
2	764	64
5	1027	94
10	1154	107
15	1176	111
20	1173	108
25	1184	114
30	1177	112
50	1173	107

Table 7.1.: Total number of genes with eQTLs and number of genes with eQTLs in *trans* for different numbers of hidden PEER factors k.

Between $k = 2$ and $k = 10$, the number of eQTLs quickly increased with an increasing number of hidden factors, suggesting that there were indeed some major experimental batch effects that PEER was effectively correcting for. Beyond $k = 10$ this positive effect seemed to weaken until the number of eQTLs started decreasing again after $k = 25$. As $k = 25$ seemed to result in the highest number both of overall eQTLs as well as *trans* eQTLs, I chose this value for all further analyses. This set contained 11,022 eQTLs for 1,184 unique genes, 203 of which (114 unique genes) were located on a chromosome arm different from their associated gene.

7.2.2. eQTLs associated with inversions

Chromosomal inversions are large-scale structural changes to chromosomes that are known to result in large regions of strong LD in the genome (see Section 1.5). Thus, if a variant associated with an inversion is an eQTL, other variants associated with the inversion will also look like they are eQTLs. This could potentially complicate the interpretation of my results, as a gene with a proximal eQTL may

have other variants in LD much further away, giving the impression that there is a distal eQTL.

There are 16 inversions that have been identified among the DGRP lines (Huang et al., 2014). I obtained the inversion genotypes for each of the DGRP lines used in my study and encoded them as minor allele dosages, with numbers from 0 to 2 indicating the number of minor alleles present in the line. For each inversion, I calculated the squared Pearson's correlation coefficient r^2 between the dosage of each inversion and each of the 1,772,891 tested variants in the genome. The distribution of the strongest r^2 per variant is shown in Figure 7.1.

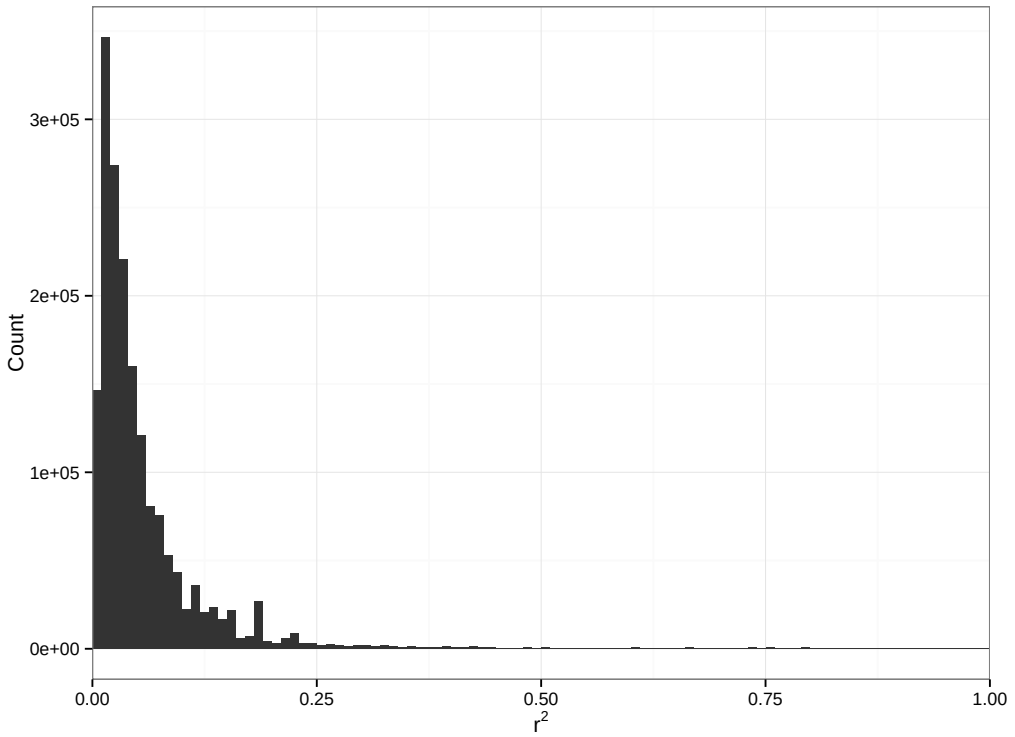
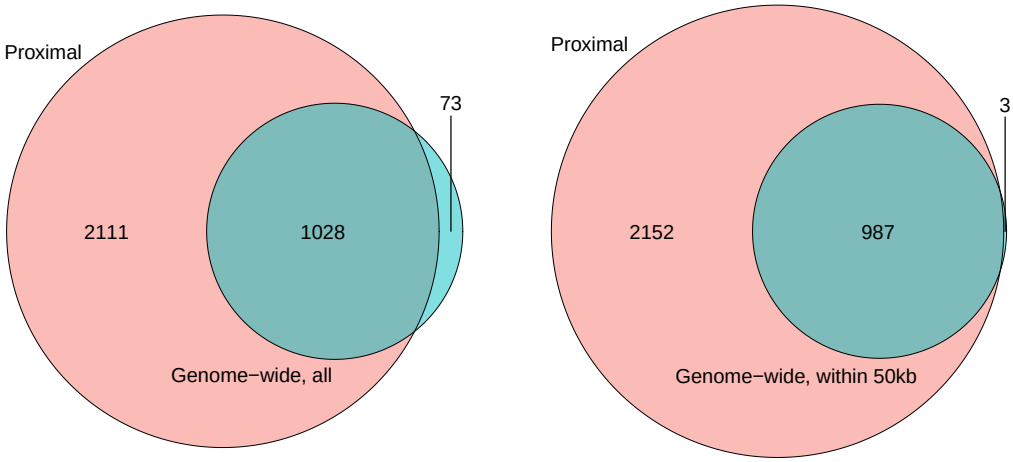


Figure 7.1.: Histogram of the strongest r^2 to an inversion for all 1,772,891 tested variants.

While most of the tested variants were not associated with any inversion, there were some that showed evidence of linkage to the inversion. As a conservative filter, I removed all eQTLs with $r^2 > 0.1$ from the analysis. After this filter, I was left with 10,301 eQTLs for 1,101 unique genes (157 eQTLs, 92 genes in *trans*).

7.3. Comparison between genome-wide and gene-proximal eQTLs

I compared this unfiltered set of 10,301 common, genome-wide eQTLs to the unfiltered set of 49,649 common, gene-proximal eQTLs described in Section 4.6. There were 2,111 genes that only had an eQTL in the proximal eQTL set, 73 genes with eQTLs only in the genome-wide eQTL set and 1,028 genes for which I found at least one eQTL with both approaches (Figure 7.2a). In fact, for 986 (96 %) of the latter genes I found exactly the same variant as an eQTL with both approaches. This overlap further increased when I only considered those variants that were tested in both analyses, namely those within 50 kb of the gene. In this case, I found 2,152 genes with eQTLs only in the proximal search, 3 genes with eQTLs only in the genome-wide search and 987 genes with both proximal and genome-wide eQTLs (Figure 7.2b).



(a) All eQTLs **(b)** Only eQTLs within 50 kb

Figure 7.2.: Overlap between genes with proximal eQTLs (red) and genome-wide eQTLs (blue) with a common effect.

Of course, this large agreement between proximal and genome-wide eQTL calls did not come as a surprise as I had applied the same model to the same data. Nevertheless, this shows that there is no large disagreement between the two different normalisation and p-value adjustment methods used. There were three genes for which I found proximal eQTLs with the genome-wide but not with the proximal test: *CG15213*, *CG31102*, and *CadN2*. In theory, I should have found these associations also during the proximal eQTL testing, since I tested the

same genes with the same variants. However, given the differences between the approaches (including running PEER with a different number of hidden factors and correcting for multiple testing in different ways) a small discrepancy like this would be expected.

The results from this overlap also allowed me to get an impression of how much larger the power to detect eQTLs was when I only tested variants in a small window around the gene instead of genome-wide. Based on the number of genes for which I found eQTLs with the two different approaches, I achieved at least a threefold increase in power by restricting the analysis to variants proximal to the gene and thus decreasing the multiple-testing burden.

7.4. Filtering and RNA-seq validation of genome-wide eQTLs

To account for mappability, I applied the same set of filtering criteria that I had applied to the proximal eQTLs (see Chapter 5) to all the genome-wide eQTLs that were not located in *trans*. If the most strongly associated variant of a gene failed these filters, I removed all other variants associated with that gene as well. After removing all eQTLs involving a region with heterozygosity $> 40\%$, eQTLs correlated with the mappability of the 3′ Tag-Seq peaks and variants closer than 25 bp to an associated 3′ Tag-Seq peak, the final set of genome-wide eQTLs contained 3,553 eQTLs for 436 genes (142 eQTLs, 86 genes in *trans*).

Next, I performed RNA-seq validation of this set of filtered genome-wide eQTLs, using the same approach as described in Section 5.1. I only tested the strongest eQTL for each gene, but giving priority to eQTLs on different chromosome arms. The concordance rate I observed in this set of 436 lead genome-wide eQTLs was 85 % (Figure 7.3). The concordance was slightly lower for those eQTLs on a different chromosome arm from their gene, but still clearly above the random expectation of 50 % (Figure 7.4).

These results confirmed that the genome-wide 3′ Tag-Seq eQTLs were unlikely to have been caused by protocol-specific biases. Again, this was not unexpected as false positives caused by mappability problems would have to be located close to the gene, so there was no reason to expect that whole-genome eQTLs would be more likely to be artefacts than proximal eQTLs. For 146 lead genome-wide eQTLs I could also see a significant difference in RNA-seq expression levels between the major and minor allele that had been predicted by the eQTL (Wilcoxon

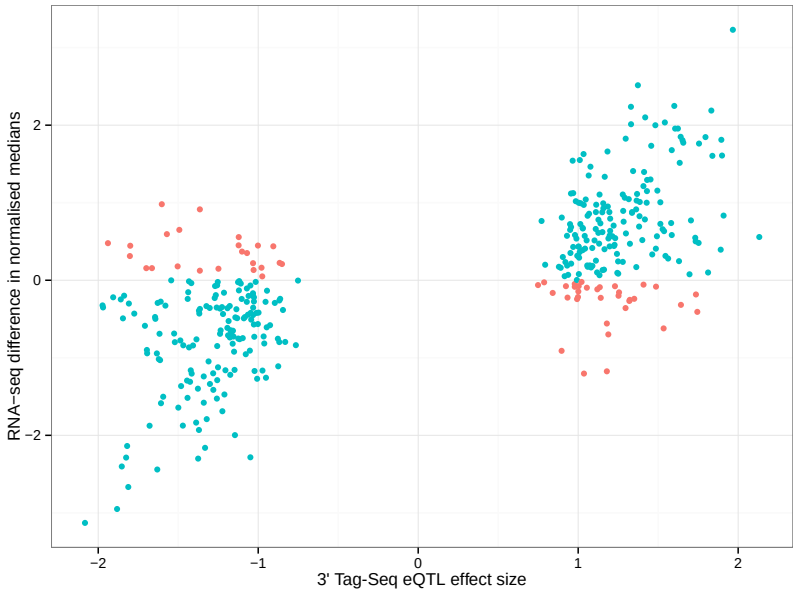


Figure 7.3.: Difference in normalised medians for expression levels measured with standard RNA-seq at 10–12 h plotted against effect size at 10–12 h of genome-wide 3′ Tag-Seq eQTLs. Concordant eQTLs shown in blue, discordant eQTLs shown in red.

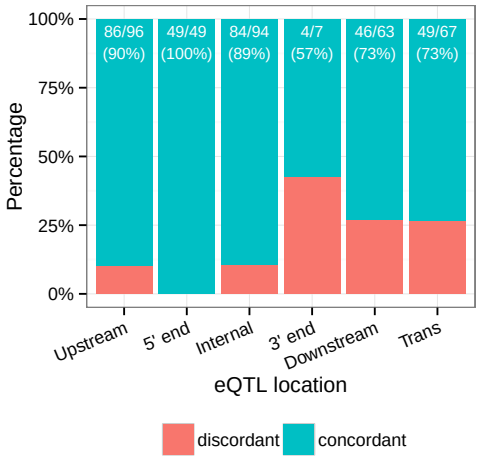


Figure 7.4.: Fraction of concordant and discordant eQTLs grouped by their location relative to their associated gene.

test, $p < 0.05$). Six of these eQTLs were located in *trans* and are shown in Table 7.2, annotated with the gene located closest to the variant and the effect of the variant based on its location and gene annotation, as predicted by the tool

snpEff (Cingolani et al., 2012). A Manhattan plot of all tested variants for the gene *CG13827*, which is a component of the peroxisome membrane (Faust et al., 2012), is shown in Figure 7.5.

Chr.	Gene	Lead variant	MAF	β	Variant effect	Closest gene
X	Traf-like	3R:14089716	0.07	-1.70	intergenic region	CG18599
3R	Spn85F	3L:13294092	0.06	1.65	upstream	Acp70A
3R	Acf1	2L:16008623	0.28	1.00	intronic	beat-Ic
3R	CG13827	2R:17778487	0.13	1.44	intronic	Fili
3R	lig3	2R:18052757	0.06	1.63	upstream	a
2L	zf30C	3R:25519022	0.11	1.70	intronic	dmrt99B

Table 7.2.: *Trans* eQTLs validated using RNA-seq expression levels. Chr., chromosome of gene with expression level change. β, effect size. Closest gene, gene located closest to the eQTL.

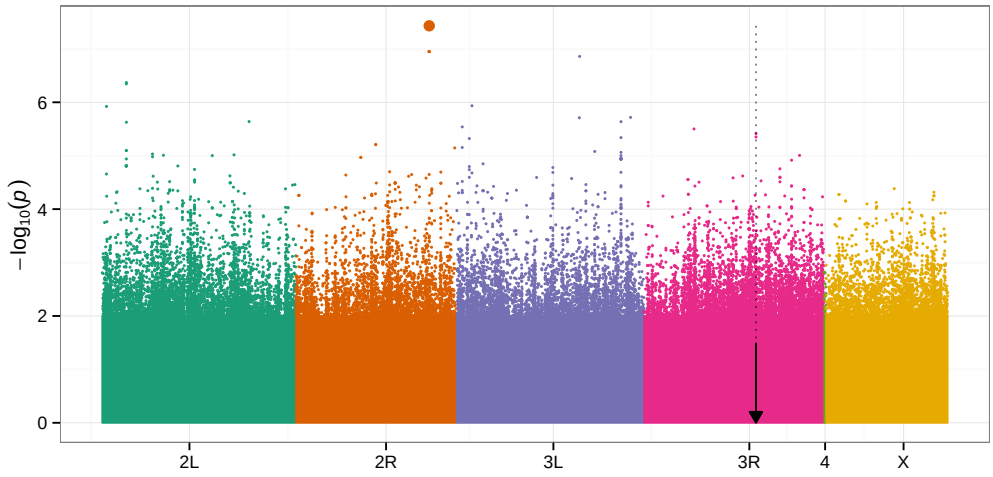


Figure 7.5.: Genome-wide Manhattan plot for the gene *CG13827*. $-\log_{10}(p)$ value of all tested variants plotted against the rank of their location, coloured by chromosome arm. Only variants with $-\log_{10}(p) > 2$ are shown as individual points, most strongly associated variant is plotted with increased size. Arrow: Location of *CG13827*.

7.5. Location of genome-wide eQTLs with respect to their genes

In Chapter 4, I made the assumption that most eQTLs strong enough to be detected would be located within 50 kb of the gene. This new genome-wide analysis now gave me the chance to test this assumption. I calculated the genomic distance of each eQTL from its associated gene, assigning an infinite distance to any eQTL on a different chromosome arm. Figure 7.6 shows the distribution of these values, truncated at 200 kb.

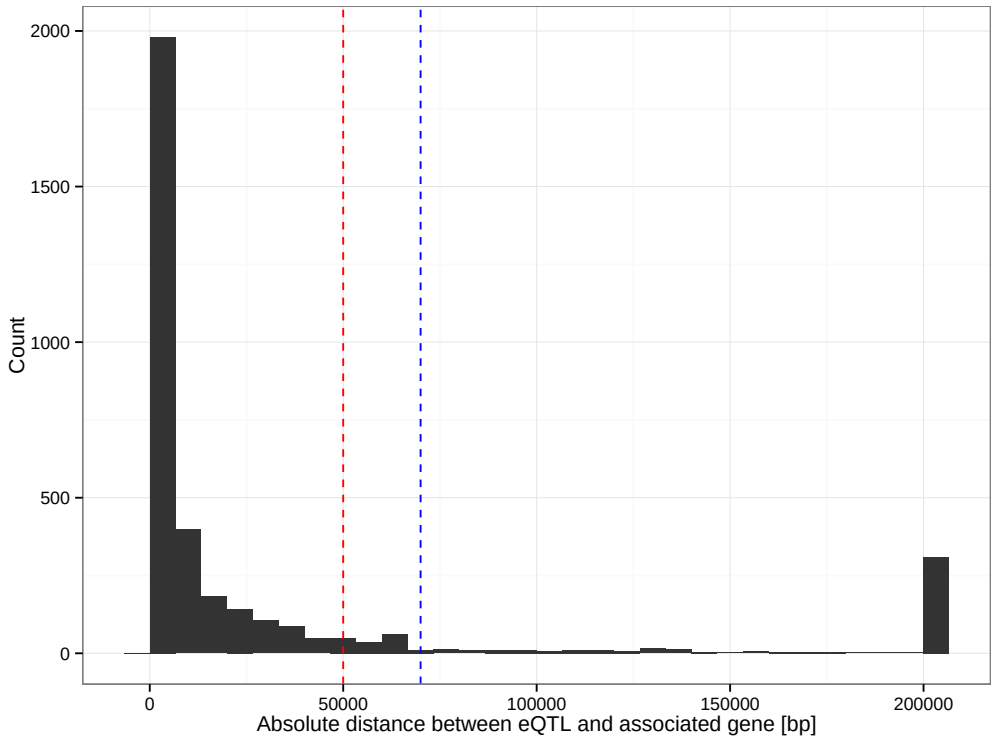


Figure 7.6.: Histogram of genome-wide eQTLs distances from their associated genes. Red line shows the 50 kb threshold used for the proximal eQTL calling, blue line another possible threshold at 70 kb. eQTLs further than 200 kb or located on a different chromosome arm from their associated gene are displayed as 200 kb.

While this distance distribution shows that eQTLs can be located many hundreds of kb away from their associated gene, the vast majority (84 %) of them were located within 50 kb (red line). This shows that 50 kb was a good choice of threshold, which allowed me to achieve good statistical power while still being

able to identify most eQTLs. However, in hindsight it might have been better to slightly extend the cut-off to 70 kb (blue line), which might have allowed me to find a few more eQTLs (87 %).

As in Section 6.1.3, I performed a GO enrichment analysis on the 115 genes that were located further than 10 kb from the TSS of their associated gene but on the same chromosome arm. I observed similar enrichments as I had already observed for the gene-proximal set of eQTLs, including the terms "single-organism developmental process", "organ development", and "response to stimulus". However, due to the lower number of genes available for this analysis in the genome-wide eQTL set, these terms no longer passed a Bonferroni-adjusted p-value threshold of 0.05, with adjusted p-values of 0.06, 0.07 and 0.08 respectively. This demonstrates that, while the genome-wide eQTL search was able to identify some distal eQTLs that were not tested in the proximal eQTL search, the increased number of eQTLs in the proximal analysis provided me with greater power to detect patterns, even at intermediate distances.

In addition, I also performed a GO enrichment analysis of the set of 106 genes that had at least one eQTL located further than 1 Mb away from the gene or on a different chromosome arm. These genes were weakly enriched in the GO terms "cell-cell signaling" and "regulation of RNA metabolic process", but the sample size was too small to be able to consider this enrichment significant after Bonferroni correction. Comparing the p-value of these terms to an empirical null from 1,000 random permutation experiments resulted in empirical p-values of 0.17 and 0.70, respectively. A list of the eight genes in the category "cell-cell signaling" with a distal eQTL is shown in Table 7.3 and a genome-wide Manhattan plot for the gene *Sra-1*, which plays a role in the development of axons (Bogdan et al., 2004), is shown in Figure 7.7. While the power to detect features of such distal eQTLs was clearly very limited, this shows that interesting associations might be uncovered in a more highly powered study.

Figure 7.8 shows the position of the 3,553 filtered genome-wide eQTLs against the position of the gene that they were associated with.

This plot confirms that most eQTLs appeared to be located very close to their associated gene but also shows a number of eQTLs further away from their gene or in *trans*. However, there did not appear to be any eQTLs affecting the expression of many different genes in *trans* at the same time. Such an eQTL would appear as a vertical line of dots in this plot, indicating that many genes are associated with the same eQTL hotspot (Albert and Kruglyak, 2015).

Chr.	Gene	Lead variant	MAF	β	Variant effect	Closest gene
3R	CG13827	2R:17778487	0.13	1.44	intronic	Fili
3R	CG31122	3R:11593798	0.20	1.02	upstream	CG5623
3L	Cip4	2R:7061368	0.25	-1.07	5′ UTR variant	shn
2L	Pde11	X:17581143	0.31	-0.97	intronic	chas
3R	Snap24	2R:18721687	0.07	-2.01	synonymous	CG30265
2R	Sply	X:22220978	0.29	-0.88	intergenic region	
3R	Sra-1	X:2375360	0.24	-0.94	splice region	trol
3L	dlt	3L:16565368	0.14	-1.84	splice region	Mipp1

Table 7.3.: Genes annotated with the GO term "cell-cell signaling" and a distal eQTL at least 1 Mb away or on a different chromosome arm. Chr., chromosome of gene with expression level change. β, effect size. Closest gene, gene located closest to the eQTL.

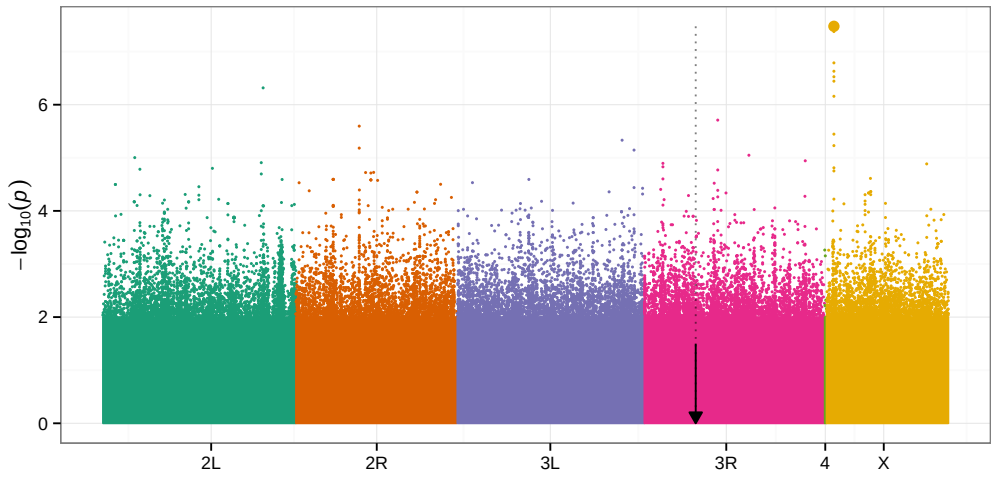


Figure 7.7.: Genome-wide Manhattan plot for the gene *Sra-1*. $-\log_{10}(p)$ value of all tested variants plotted against the rank of their location, coloured by chromosome arm. Only variants with $-\log_{10}(p) > 2$ are shown as individual points, most strongly associated variant is plotted with increased size. Arrow: Location of *Sra-1*.

Finally, I searched for *trans* eQTLs that were predicted to affect the protein coding region of a gene. These cases were particularly interesting, as they could have involved polymorphisms causing a structural change to a transcriptional regulator, which then resulted in the decreased binding of that regulator to a binding site. There were two such eQTLs, shown in table Table 7.4.

The first eQTL had a negative effect on the expression level of the gene *CG-12096*, which is annotated with the GO terms "cellular response to DNA damage

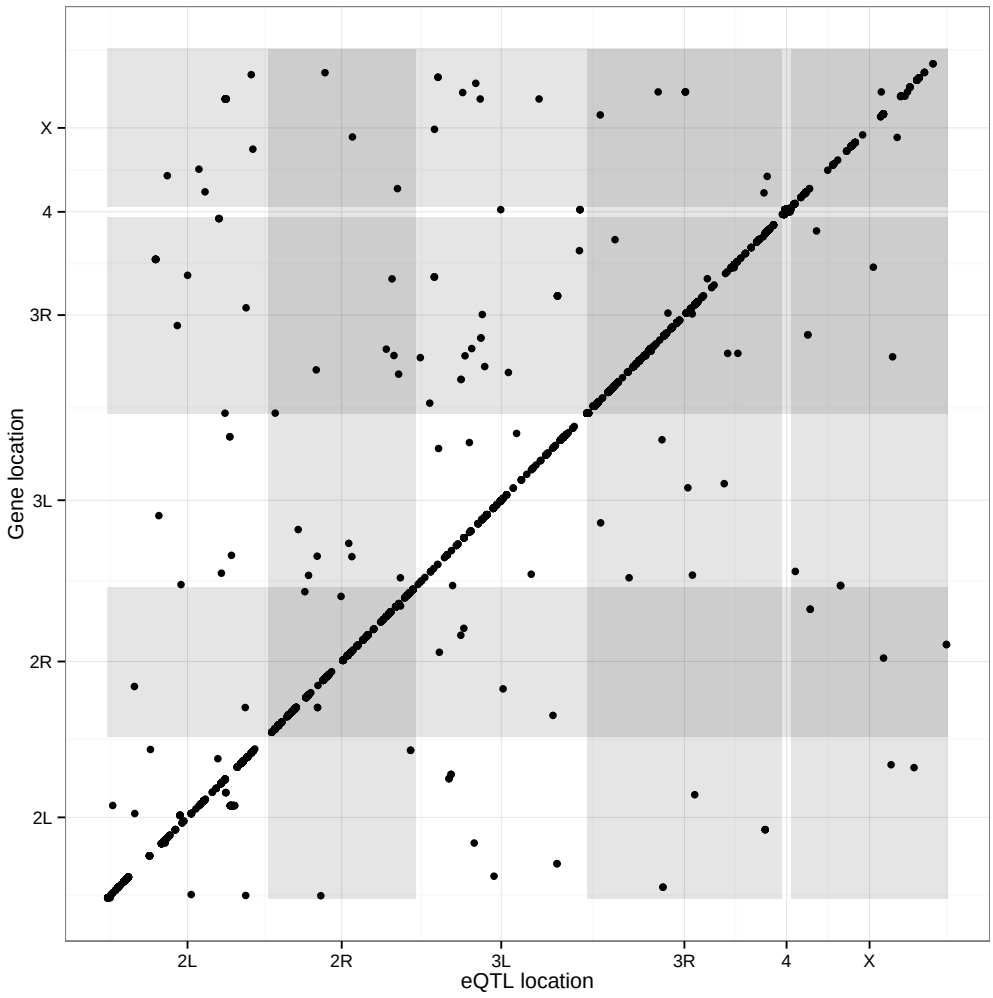


Figure 7.8.: Location of genes (Y-axis) plotted against the location of their association variant (X-axis) for all 3,553 filtered genome-wide eQTLs. Chromosome arms shaded in alternating grey/white pattern.

stimulus" (Ravi et al., 2009) and "proteasome assembly" (Cho-Park and Steller, 2013). The protein affected by the variant itself was Or83c, which is a Odorant receptor (Ronderos et al., 2014). Its 31st amino acid, Serine, is changed to Alanine due to a T to G polymorphism at the variant. The second eQTL decreased the expression of *CG10289*. This gene is not annotated with any GO terms, but has been associated with the PpV phosphatase (Yin et al., 2014). The variant changes the 38th amino acid in the gene *CG30087*, which has been annotated with "serine-type endopeptidase activity", from Threonine to Proline through an

Chr.	Gene	Lead variant	MAF	β	Variant effect	Closest gene
X	CG12096	3R:1903915	0.12	-1.33	missense variant	Or83c
3L	CG10289	2R:11580064	0.30	-1.07	missense variant	CG30087

Table 7.4.: eQTLs that are predicted to have a non-synonymous effect on a gene in *trans*. Chr., chromosome of gene with expression level change. β, effect size.

A to C polymorphism. Neither of these cases thus appeared to have an obvious mechanistic explanation, which may mean that they were false positives but could also make them interesting targets for further study.

The Manhattan plot for *CG12096* is shown in Figure 7.9; the Manhattan plot for *CG10289* is shown in Figure 7.10.

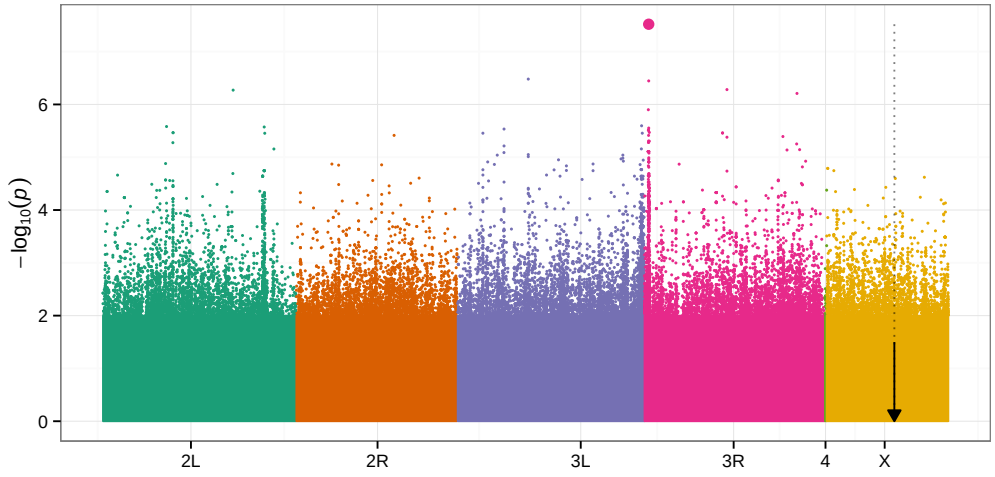


Figure 7.9.: Genome-wide Manhattan plot for the gene *CG12096*. $-\log_{10}(p)$ value of all tested variants plotted against the rank of their location, coloured by chromosome arm. Only variants with $-\log_{10}(p) > 2$ are shown as individual points, most strongly associated variant is plotted with increased size. Arrow: Location of *CG12096*.

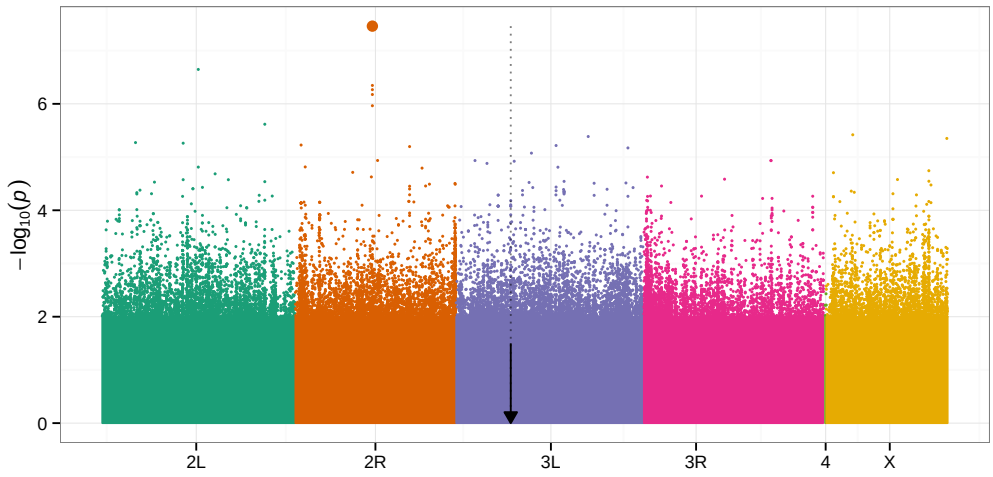


Figure 7.10.: Genome-wide Manhattan plot for the gene *CG10289*. $-\log_{10}(p)$ value of all tested variants plotted against the rank of their location, coloured by chromosome arm. Only variants with $-\log_{10}(p) > 2$ are shown as individual points, most strongly associated variant is plotted with increased size. Arrow: Location of *CG10289*.

7.6. Comparison to variance decomposition

The results from the variance decomposition analysis (see Section 4.2) suggested that *trans* regulation was prevalent in the *D. melanogaster* genome, with 15 % of genes having more than 25 % of their total variance in expression level explained by distal genetic relatedness (putative *trans* effects). The reason that I was nevertheless unable to identify *trans* eQTLs with an effect on many genes is thus unlikely to be reflective of a lack of real *trans* regulation, and may instead have been caused by some or all of the following factors:

1. Low power: Correcting for multiple testing across all variants in the genome vastly decreased the statistical power. Additionally, *trans* effects tend to have low effect sizes, which makes them hard to detect (Petretto et al., 2006). Since I only used 228 samples from 76 different lines for my analysis, I was unlikely to detect any such weak effects.

2. Selection: A mutation changing the expression level of many genes at the same time in *trans* is likely to have a more extreme effect than a single *cis* regulatory mutation, and may thus be under stronger selection (Wray, 2007). It is possible that such effects would not be present in the population

at my MAF threshold of 5 %.

3. Polygenic effects: The *trans* effects that I observed in the variance decomposition may have actually been the sum of many small effects, all acting in *trans*. I would have been able to detect such effects in the variance decomposition since I was considering all distal variants at the same time, but the effects of the individual variants may have been below the threshold of detectable effects in this study.

8. Concluding remarks

In this work, I have described an eQTL study in the *Drosophila* Genetic Reference Panel, which has yielded interesting new insights into gene regulation. Thanks to the high degree of fine mapping possible in this population, I was often able to identify the location of individual nucleotides associated with gene expression and study their exact location with respect to the gene. I found eQTLs proximal to the promoter, but also many around the 3′ end of the gene or in more distal regions. In addition, the nucleotide-level accuracy of this study allowed me to find eQTLs affecting potential novel binding sites involved in gene regulation.

Using the 3′ isoform expression data generated by 3′ Tag-Seq, I could not only annotate new 3′ polyadenylation sites genome-wide, but also identify a new type of QTL affecting alternative polyadenylation. These apaQTLs were associated with a switch in 3′ polyadenylation site usage, both with and without a change in the total gene expression level.

The multi-stage experimental design further enhanced this study, allowing me to search for gene regulatory differences between different stages of embryo development. This revealed some interesting stage-specific effects that illustrate the plasticity of gene regulation throughout development. My observation that eQTLs with a strong effect at 2–4 h after fertilisation tended to be enriched in exons suggests that post-transcriptional regulatory mechanisms may be more prevalent at early stages of development.

In addition, I have shown how eQTL and gene expression data can be employed in the study of organismal phenotypes, using the example of *Orct2*. Such an approach could be used to help our understanding of the causal relationships between a phenotype and the expression levels of associated genes in a process similar to Mendelian randomisation, which has been used to identify causal associations in human health (Lawlor et al., 2008).

Finally, I have developed a large collection of data sets in the course of this project, which I hope will prove useful for the entire *Drosophila* community. Using the data sets that we have submitted for publication (Cannavo et al., 2015),

other researchers will be able to find alternative polyadenylation sites, compare expression levels, and identify regulatory regions for thousands of genes in *D. melanogaster*.

8.1. Possible improvements to this study

In hindsight, there were of course some aspects of this study that could have been improved. First and foremost, an increase in the sample size would not only have increased my power to find more proximal eQTLs, but would also have allowed me to conduct a full study of *trans*-acting eQTLs, which was underpowered at the current sample size. Samples from additional developmental stages would have also been very helpful in disentangling common from stage-specific effects. Similarly, a fully randomised sample preparation and sequencing design would have made it easier to limit batch effects, requiring fewer normalisation steps.

The data generated by 3′ Tag-Seq also proved to be challenging to process and analyse, as most of the commonly used tools had been designed for standard RNA-seq. The poly(A)$^+$ RNA-seq protocol may have thus been a better choice to estimate gene expression levels, and may have also been less susceptible to mappability artefacts. However, this would have meant that I would not have been able to identify apaQTLs or study 3′ transcript end locations in this amount of detail.

In addition to these changes to the experimental design, there are also some improvements that I could have made to the way I analysed the data. If I had filtered the variants to be tested more aggressively and removed variants of low quality or in strong LD to another variant, I could have potentially decreased the number of tests I needed to perform, without missing any positive associations. This could have increased my power, particularly in the genome-wide eQTL analysis. A more carefully chosen mappability filter might have also allowed me to remove more false positives while keeping more true positives.

8.2. Future steps

I believe that there are many interesting analyses that could be performed to extend and build upon this work.

On the computational side, the data I generated represents a treasure trove of information that could be very useful for studies of gene expression and gene

regulation in *Drosophila*. For example, in 2009, Ayroles and colleagues studied the systems genetics of gene expression in 40 *Drosophila* Genetic Reference Panel lines using microarray data from adult individuals (Ayroles et al., 2009). Using the data from my project, this study could now be repeated with a doubled sample size, a more accurate method of determining gene expression levels and the additional dimension of multiple developmental time points.

In addition, my approach to the developmental staging of *Drosophila* by gene expression levels could be extended much further. Using machine learning techniques it may even be possible to predict the exact developmental time point of each individual. This would allow one to consider the developmental time point as a covariate in the eQTL study, which could yield very interesting results. The concept of determining the developmental stage of an organism from its gene expression levels could also be expanded to other organisms.

It would also be very interesting to conduct additional *in vitro* and *in vivo* assays to follow up on the phenomena that I observed, to help confirm some of my results and study their impact on the organism. For example, it remains to be seen what effects an experimental disruption of the novel binding site motifs that I found would have on the expression level of genes. In addition, the effect of eQTLs on organismal phenotypes could be studied further, as I have shown with the example of *Orct2*. This could involve both the collection of additional phenotyping data as well as the targeted modification of the DNA at a given eQTL, using a technique such as the *CRISPR/Cas9* system (Bassett et al., 2013; Hsu et al., 2014).

A heterozygous F_1 cross of different lines from the DGRP could also be an interesting resource for further study. For example, these offspring could be tested for allele-specific expression, which would make it possible to not only validate the effect of some of the eQTLs that I have identified, but also distinguish *cis* and *trans* regulatory effects in a much more accurate manner. Conducting multiple such crosses between different lines and analysing gene expression at multiple stages of development might yield completely new insights into the heredity and plasticity of gene regulation.

Recently, single-cell RNA-seq has also become available, enabling the measurement of gene expression levels in individual cells (Tang et al., 2009). While this technology is still being developed, recent improvements now allow for the analysis of gene expression levels of thousands of individual cells at the same time (Macosko et al., 2015). This may make it possible to study the transcriptome

of different cell types in the developing *Drosophila* embryo to a level that is not feasible with the current whole-embryo RNA-seq data. Obtaining this data from multiple individuals could even allow for an eQTL study based on single-cell sequencing data, although the cost of this would most likely be prohibitive at the current time.

8.3. Genetics on the fly

Since Morgan started his work with *Drosophila melanogaster* more than a century ago, the fruit fly has played a major (and arguably the most important) role in the study of genetics and development. Yet in spite of these decades of close examination, there are still many questions left unanswered and new experimental methods continuously open up new avenues of study. In this thesis, I have contributed to our understanding of gene regulation by using genetic association techniques in this model organism. There are many other molecular and organismal phenotypes in *Drosophila* to which this technique could be applied. Thus, I believe that the study of *Drosophila melanogaster* will continue to bring exciting new discoveries in the decades and centuries to come.

A. Supplementary Table

A.1. Samples used in this study

Time	Line	Repl.	Run	Lane	Barcode	Reads	QC	eQTLs
2–4 h	21	1	22	C1RHM	TTGCGG	11781844	✓	✓
6–8 h	21	1	7	D18BD	TCCGTC	12443917	✓	✓
10–12 h	21	1	12	C1FY9	AAGTGC	11765316	✓	✓
2–4 h	26	1	18	C1J9A	GATGCT	16020124	✓	✓
6–8 h	26	1	10	C1CKM	GATGCT	15509549	✓	✓
10–12 h	26	1	17	C1J9A	GATGCT	16928923	✓	✓
2–4 h	28	1	22	C1RHM	CGTACG	15305686	✓	✓
2–4 h	28	2	24	C1TAU	CCGTAT	11231565	✓	–
6–8 h	28	1	10	C1CKM	CCACTC	15783639	✓	✓
6–8 h	28	2	24	C1TAU	TTGCGG	11361119	–	–
10–12 h	28	1	17	C1J9A	CCGTAT	13711981	–	–
10–12 h	28	2	23	C1TAU	GATGCT	12888337	✓	✓
2–4 h	40	1	22	C1RHM	TACAAG	13173984	✓	✓
2–4 h	40	2	24	C1TAU	CGTACG	10826737	✓	–
6–8 h	40	1	24	C1TAU	TACAAG	11817078	✓	✓
10–12 h	40	1	24	C1TAU	TCCGTC	11280581	✓	✓
2–4 h	41	1	21	C1J9A	GATGCT	21951300	✓	✓
6–8 h	41	1	10	C1CKM	AAGTGC	15306165	✓	✓
10–12 h	41	1	25	C258F	GATGCT	15466033	✓	✓
2–4 h	42	1	18	C1J9A	CCGTAT	12453968	✓	✓
6–8 h	42	1	8	C1CKM	GATGCT	15649897	✓	✓
10–12 h	42	1	14	C1FY9	GATGCT	14773016	✓	✓
2–4 h	57	1	20	C1J9A	GATGCT	14301333	✓	✓
6–8 h	57	1	8	C1CKM	CCGTAT	12824090	✓	✓
10–12 h	57	1	13	C1FY9	TACAAG	15722406	✓	✓

Time	Line	Repl.	Run	Lane	Barcode	Reads	QC	eQTLs
2–4 h	75	1	22	C1RHM	TCCGTC	13256971	✓	✓
6–8 h	75	1	10	C1CKM	ATTATA	14548749	✓	✓
10–12 h	75	1	17	C1J9A	TTGCGG	11993024	✓	✓
2–4 h	83	1	24	C1TAU	CCACTC	13005031	✓	✓
6–8 h	83	1	23	C1TAU	CCGTAT	11664281	✓	✓
10–12 h	83	1	23	C1TAU	TTGCGG	12447528	✓	✓
2–4 h	93	1	22	C1RHM	GATGCT	16337643	✓	✓
6–8 h	93	1	10	C1CKM	GGAGAA	16053679	✓	✓
10–12 h	93	1	14	C1FY9	CCGTAT	14606457	✓	✓
2–4 h	129	1	21	C1J9A	CCGTAT	12186219	✓	✓
6–8 h	129	1	10	C1CKM	CCGTAT	13862381	✓	✓
10–12 h	129	1	17	C1J9A	CGTACG	13762668	✓	✓
2–4 h	176	1	22	C1RHM	CCGTAT	11394736	✓	✓
6–8 h	176	1	8	C1CKM	TTGCGG	16844815	✓	✓
10–12 h	176	1	13	C1FY9	TTGCGG	13057662	✓	✓
2–4 h	177	1	20	C1J9A	CCGTAT	13912216	✓	✓
6–8 h	177	1	7	D18BD	CCACTC	11817487	✓	✓
6–8 h	177	2	8	C1CKM	CGTACG	14263778	✓	–
10–12 h	177	1	25	C258F	CCGTAT	13881302	✓	✓
2–4 h	181	1	20	C1J9A	TTGCGG	11653877	✓	✓
6–8 h	181	1	8	C1CKM	TACAAG	11989556	✓	✓
10–12 h	181	1	12	C1FY9	ATTATA	11893302	✓	✓
2–4 h	208	1	16	D1MWE	CCGTAT	11889102	✓	✓
6–8 h	208	1	5	C0PPA	TCCGTC	13926767	✓	✓
10–12 h	208	1	11	C1CKM	GATGCT	14101030	✓	✓
2–4 h	227	1	22	C1RHM	CCACTC	11872220	✓	✓
6–8 h	227	1	7	D18BD	AAGTGC	11644217	✓	✓
10–12 h	227	1	12	C1FY9	TACAAG	12941098	✓	✓
2–4 h	239	1	25	C258F	TTGCGG	12388091	✓	✓
6–8 h	239	1	12	C1FY9	GATGCT	13683224	✓	✓
10–12 h	239	1	17	C1J9A	TACAAG	15586529	✓	✓
2–4 h	280	1	21	C1J9A	TTGCGG	9480810	✓	✓
6–8 h	280	1	12	C1FY9	TTGCGG	10695199	✓	✓
10–12 h	280	1	17	C1J9A	CCACTC	14553632	✓	✓

Time	Line	Repl.	Run	Lane	Barcode	Reads	QC	eQTLs
2–4 h	301	1	15	C1J9A	CGTACG	15816037	✓	✓
6–8 h	301	1	2	D150P	TCCGTC	10914175	✓	✓
10–12 h	301	1	3	D150P	TCCGTC	12733925	✓	✓
2–4 h	303	1	15	C1J9A	TACAAG	13372145	✓	✓
6–8 h	303	1	4	D150P	GATGCT	14978759	✓	✓
10–12 h	303	1	6	D18BD	GATGCT	11812297	✓	✓
2–4 h	304	1	16	D1MWE	GATGCT	13513702	✓	✓
6–8 h	304	1	2	D150P	CCACTC	12806480	✓	✓
10–12 h	304	1	6	D18BD	CCGTAT	10571228	✓	✓
2–4 h	307	1	19	C1J9A	TCCGTC	13646698	✓	✓
6–8 h	307	1	3	D150P	GATGCT	13654237	✓	✓
10–12 h	307	1	9	C1CKM	GATGCT	13173636	✓	✓
2–4 h	313	1	15	C1J9A	CCGTAT	14048595	✓	✓
6–8 h	313	1	2	D150P	GATGCT	12001999	–	–
6–8 h	313	2	5	C0PPA	CCACTC	11060246	✓	✓
10–12 h	313	1	6	D18BD	TTGCGG	8361208	✓	✓
2–4 h	318	1	18	C1J9A	TTGCGG	14207627	✓	✓
6–8 h	318	1	10	C1CKM	TTGCGG	14145307	✓	✓
10–12 h	318	1	19	C1J9A	GATGCT	13681324	✓	✓
2–4 h	320	1	18	C1J9A	CGTACG	14116237	✓	✓
6–8 h	320	1	11	C1CKM	CCACTC	13723436	✓	✓
10–12 h	320	1	13	C1FY9	CCGTAT	13209356	✓	✓
2–4 h	324	1	16	D1MWE	TTGCGG	10853915	✓	✓
6–8 h	324	1	5	C0PPA	GGAGAA	13775191	✓	✓
10–12 h	324	1	11	C1CKM	CCGTAT	14837755	✓	✓
2–4 h	335	1	25	C258F	ATTATA	12638023	✓	✓
6–8 h	335	1	25	C258F	GGAGAA	11933184	✓	✓
10–12 h	335	1	26	C258F	AAGTGC	13070336	✓	✓
2–4 h	357	1	19	C1J9A	CCACTC	14660601	✓	✓
6–8 h	357	1	3	D150P	CCGTAT	13852505	✓	✓
10–12 h	357	1	9	C1CKM	CCGTAT	14690063	✓	✓
2–4 h	358	1	16	D1MWE	CGTACG	12369969	✓	✓
6–8 h	358	1	5	C0PPA	CGTACG	11668578	✓	✓
10–12 h	358	1	9	C1CKM	TTGCGG	10810663	✓	✓

Time	Line	Repl.	Run	Lane	Barcode	Reads	QC	eQTLs
2–4 h	360	1	16	D1MWE	TACAAG	13654383	✓	✓
6–8 h	360	1	5	C0PPA	TACAAG	17035971	✓	✓
10–12 h	360	1	11	C1CKM	TTGCGG	13398508	✓	✓
2–4 h	362	1	16	D1MWE	TCCGTC	12907513	✓	✓
6–8 h	362	1	5	C0PPA	ATTATA	10011283	✓	✓
10–12 h	362	1	9	C1CKM	TCCGTC	14754377	✓	✓
2–4 h	365	1	15	C1J9A	GGAGAA	13465344	✓	✓
6–8 h	365	1	2	D150P	TACAAG	10684547	✓	✓
10–12 h	365	1	6	D18BD	TACAAG	12400547	✓	✓
2–4 h	370	1	20	C1J9A	TACAAG	11915348	✓	✓
6–8 h	370	1	11	C1CKM	AAGTGC	14704146	✓	✓
10–12 h	370	1	12	C1FY9	TCCGTC	12646054	✓	✓
2–4 h	374	1	18	C1J9A	TACAAG	13922196	✓	✓
2–4 h	374	2	23	C1TAU	CGTACG	11705149	✓	–
6–8 h	374	1	24	C1TAU	AAGTGC	11902037	✓	✓
10–12 h	374	1	19	C1J9A	CCGTAT	11983094	✓	✓
10–12 h	374	2	23	C1TAU	TACAAG	13172663	✓	–
2–4 h	375	1	14	C1FY9	CCACTC	13132406	✓	✓
6–8 h	375	1	1	C0R6L	TTGCGG	8039426	✓	✓
6–8 h	375	2	1	C0R6L	CGTACG	10016269	✓	–
6–8 h	375	3	1	C0R6L	TACAAG	7930105	✓	–
6–8 h	375	4	1	C0R6L	AAGTGC	13628433	✓	–
6–8 h	375	5	1	C0R6L	TCCGTC	12677795	✓	–
10–12 h	375	1	3	D150P	TTGCGG	11771741	✓	✓
2–4 h	379	1	15	C1J9A	TTGCGG	13021221	✓	✓
6–8 h	379	1	5	C0PPA	TTGCGG	12754934	✓	✓
10–12 h	379	1	11	C1CKM	CGTACG	14690477	✓	✓
2–4 h	380	1	16	D1MWE	AAGTGC	12756991	✓	✓
6–8 h	380	1	4	D150P	TACAAG	16415155	✓	✓
10–12 h	380	1	6	D18BD	TCCGTC	12284806	✓	✓
2–4 h	391	1	19	C1J9A	AAGTGC	14965944	✓	–
10–12 h	391	1	9	C1CKM	CCACTC	14910779	✓	–
2–4 h	399	1	15	C1J9A	TCCGTC	11556094	✓	✓
6–8 h	399	1	5	C0PPA	GATGCT	13796695	✓	✓

Time	Line	Repl.	Run	Lane	Barcode	Reads	QC	eQTLs
10–12 h	399	1	9	C1CKM	AAGTGC	15003665	✓	✓
2–4 h	406	1	18	C1J9A	TCCGTC	14880631	✓	✓
6–8 h	406	1	7	D18BD	GATGCT	11352918	✓	✓
10–12 h	406	1	24	C1TAU	GATGCT	14456415	✓	✓
2–4 h	427	1	16	D1MWE	ATTATA	13701215	✓	✓
6–8 h	427	1	1	C0R6L	CCACTC	9019205	✓	✓
10–12 h	427	1	6	D18BD	CCACTC	12115891	✓	✓
2–4 h	437	1	15	C1J9A	CCACTC	14630484	✓	✓
6–8 h	437	1	2	D150P	CCGTAT	10915500	✓	✓
10–12 h	437	1	6	D18BD	AAGTGC	11596118	✓	✓
2–4 h	441	1	21	C1J9A	CGTACG	8826935	✓	✓
6–8 h	441	1	11	C1CKM	ATTATA	14731230	✓	✓
10–12 h	441	1	17	C1J9A	AAGTGC	15687845	✓	✓
2–4 h	461	1	20	C1J9A	CCACTC	13338891	✓	✓
6–8 h	461	1	7	D18BD	CCGTAT	10883516	✓	✓
10–12 h	461	1	13	C1FY9	GATGCT	14248775	✓	✓
2–4 h	486	1	19	C1J9A	ATTATA	14320964	✓	✓
6–8 h	486	1	4	D150P	CCACTC	15438102	✓	✓
6–8 h	486	2	4	D150P	TCCGTC	14131598	✓	–
10–12 h	486	1	9	C1CKM	ATTATA	19097360	✓	✓
2–4 h	491	1	20	C1J9A	TCCGTC	11168114	✓	✓
6–8 h	491	1	7	D18BD	ATTATA	10113031	✓	✓
10–12 h	491	1	24	C1TAU	ATTATA	11615487	✓	✓
2–4 h	508	1	23	C1TAU	TCCGTC	11745552	✓	✓
6–8 h	508	1	23	C1TAU	CCACTC	13030804	✓	✓
10–12 h	508	1	24	C1TAU	GGAGAA	10275782	✓	✓
2–4 h	509	1	25	C258F	TCCGTC	13000929	–	–
6–8 h	509	1	26	C258F	TTGCGG	10915133	✓	–
10–12 h	509	1	25	C258F	TACAAG	11894923	–	–
2–4 h	517	1	14	C1FY9	GGAGAA	13865949	✓	✓
6–8 h	517	1	4	D150P	CCGTAT	14657885	✓	✓
10–12 h	517	1	3	D150P	AAGTGC	13564087	✓	✓
2–4 h	531	1	26	C258F	TCCGTC	13241308	✓	✓
6–8 h	531	1	26	C258F	CCGTAT	12386145	✓	✓

Time	Line	Repl.	Run	Lane	Barcode	Reads	QC	eQTLs
10–12 h	531	1	25	C258F	CCACTC	11289089	✓	✓
2–4 h	535	1	20	C1J9A	AAGTGC	16087068	✓	✓
6–8 h	535	1	7	D18BD	GGAGAA	10541169	✓	✓
10–12 h	535	1	25	C258F	CGTACG	14911611	✓	✓
2–4 h	555	1	20	C1J9A	ATTATA	13533782	✓	✓
6–8 h	555	1	4	D150P	AAGTGC	16109231	✓	✓
10–12 h	555	1	11	C1CKM	TCCGTC	17341619	✓	✓
2–4 h	639	1	16	D1MWE	CCACTC	11461026	✓	✓
6–8 h	639	1	5	C0PPA	AAGTGC	8638928	✓	✓
10–12 h	639	1	9	C1CKM	GGAGAA	15659478	✓	✓
2–4 h	642	1	18	C1J9A	CCACTC	14837681	✓	✓
6–8 h	642	1	10	C1CKM	CGTACG	16326137	✓	✓
10–12 h	642	1	19	C1J9A	TTGCGG	12721184	✓	✓
2–4 h	703	1	18	C1J9A	AAGTGC	15900554	✓	✓
6–8 h	703	1	8	C1CKM	CCACTC	10102112	✓	✓
10–12 h	703	1	13	C1FY9	TCCGTC	17182039	✓	✓
2–4 h	705	1	20	C1J9A	GGAGAA	12548270	✓	✓
6–8 h	705	1	4	D150P	ATTATA	15518714	✓	✓
10–12 h	705	1	9	C1CKM	TACAAG	17101670	✓	✓
2–4 h	707	1	19	C1J9A	GGAGAA	14488320	✓	✓
6–8 h	707	1	3	D150P	CGTACG	14620265	✓	✓
10–12 h	707	1	9	C1CKM	CGTACG	12746206	✓	✓
2–4 h	712	1	16	D1MWE	GGAGAA	14068272	✓	✓
6–8 h	712	1	5	C0PPA	CCGTAT	11123448	✓	✓
10–12 h	712	1	11	C1CKM	TACAAG	14793132	✓	✓
2–4 h	714	1	14	C1FY9	TCCGTC	12198641	✓	✓
6–8 h	714	1	4	D150P	CGTACG	16010824	✓	✓
10–12 h	714	1	4	D150P	GGAGAA	12633038	✓	✓
2–4 h	716	1	21	C1J9A	TACAAG	10372292	✓	✓
6–8 h	716	1	10	C1CKM	TACAAG	20292861	✓	✓
10–12 h	716	1	17	C1J9A	GGAGAA	15202525	✓	✓
2–4 h	721	1	21	C1J9A	TCCGTC	20211499	✓	✓
6–8 h	721	1	10	C1CKM	TCCGTC	14115601	✓	✓
10–12 h	721	1	19	C1J9A	CGTACG	14517882	✓	✓

Time	Line	Repl.	Run	Lane	Barcode	Reads	QC	eQTLs
2–4 h	730	1	26	C258F	CCACTC	12242783	✓	✓
6–8 h	730	1	26	C258F	GATGCT	14280745	✓	✓
10–12 h	730	1	26	C258F	TACAAG	13722762	✓	✓
2–4 h	732	1	14	C1FY9	AAGTGC	13261726	✓	✓
6–8 h	732	1	2	D150P	TTGCGG	9845147	✓	✓
10–12 h	732	1	6	D18BD	ATTATA	12075591	✓	✓
2–4 h	765	1	14	C1FY9	TTGCGG	13934616	✓	✓
6–8 h	765	1	4	D150P	TTGCGG	16308325	✓	✓
10–12 h	765	1	3	D150P	ATTATA	14285231	✓	✓
2–4 h	774	1	15	C1J9A	AAGTGC	13948525	✓	✓
6–8 h	774	1	2	D150P	ATTATA	14267311	✓	✓
10–12 h	774	1	3	D150P	GGAGAA	14355524	✓	✓
2–4 h	786	1	14	C1FY9	ATTATA	14215097	✓	✓
6–8 h	786	1	2	D150P	CGTACG	13145134	✓	✓
10–12 h	786	1	6	D18BD	GGAGAA	11896970	✓	✓
2–4 h	790	1	21	C1J9A	CCACTC	9956160	✓	✓
6–8 h	790	1	7	D18BD	TTGCGG	11497242	✓	✓
10–12 h	790	1	13	C1FY9	CCACTC	16668866	✓	✓
2–4 h	799	1	14	C1FY9	CGTACG	14924156	✓	–
6–8 h	799	1	1	C0R6L	GGAGAA	9429397	–	–
10–12 h	799	1	3	D150P	TACAAG	13781951	✓	–
2–4 h	804	1	18	C1J9A	ATTATA	13361712	✓	✓
6–8 h	804	1	12	C1FY9	CCGTAT	13495211	✓	✓
10–12 h	804	1	17	C1J9A	ATTATA	14678740	✓	✓
2–4 h	805	1	20	C1J9A	CGTACG	12253214	✓	✓
6–8 h	805	1	8	C1CKM	TCCGTC	14397449	✓	✓
10–12 h	805	1	13	C1FY9	CGTACG	14867230	✓	✓
2–4 h	810	1	23	C1TAU	AAGTGC	14121610	✓	✓
6–8 h	810	1	23	C1TAU	ATTATA	12849453	✓	✓
10–12 h	810	1	25	C258F	AAGTGC	10454738	✓	✓
2–4 h	820	1	15	C1J9A	ATTATA	12111362	✓	✓
6–8 h	820	1	2	D150P	GGAGAA	11568504	✓	✓
6–8 h	820	2	1	C0R6L	ATTATA	10667592	✓	–
10–12 h	820	1	6	D18BD	CGTACG	12491873	✓	✓

Time	Line	Repl.	Run	Lane	Barcode	Reads	QC	eQTLs
2–4 h	852	1	14	C1FY9	TACAAG	15121258	✓	–
6–8 h	852	1	1	C0R6L	GATGCT	20274449	✓	–
6–8 h	852	2	1	C0R6L	CCGTAT	15143335	✓	–
10–12 h	852	1	26	C258F	CGTACG	12514736	–	–
2–4 h	859	1	22	C1RHM	AAGTGC	15732503	✓	✓
6–8 h	859	1	7	D18BD	CGTACG	11693671	✓	✓
10–12 h	859	1	13	C1FY9	AAGTGC	14677842	✓	✓
2–4 h	879	1	21	C1J9A	AAGTGC	9999969	✓	✓
6–8 h	879	1	12	C1FY9	CCACTC	13593490	✓	✓
10–12 h	879	1	19	C1J9A	TACAAG	14409397	✓	✓
2–4 h	887	1	23	C1TAU	GGAGAA	12170735	✓	✓
6–8 h	887	1	11	C1CKM	GGAGAA	19557175	✓	✓
10–12 h	887	1	17	C1J9A	TCCGTC	14103103	✓	✓
2–4 h	890	1	18	C1J9A	GGAGAA	13352966	✓	✓
6–8 h	890	1	7	D18BD	TACAAG	13129505	✓	✓
10–12 h	890	1	12	C1FY9	CGTACG	13292262	✓	✓
2–4 h	892	1	21	C1J9A	ATTATA	11652415	✓	✓
6–8 h	892	1	8	C1CKM	AAGTGC	12842089	✓	✓
10–12 h	892	1	13	C1FY9	ATTATA	14072234	✓	✓
2–4 h	897	1	21	C1J9A	GGAGAA	9453103	✓	✓
6–8 h	897	1	8	C1CKM	ATTATA	12298790	✓	✓
10–12 h	897	1	13	C1FY9	GGAGAA	14633849	✓	✓
2–4 h	908	1	22	C1RHM	ATTATA	15158181	✓	✓
6–8 h	908	1	8	C1CKM	GGAGAA	12822587	✓	✓
10–12 h	908	1	12	C1FY9	GGAGAA	11892591	✓	✓

Table A.1.: List of all samples sequenced for this study. Repl., replicate. Run, sequencing run. Lane, sequencing lane used. Reads, number of sequenced reads. QC, did this sample pass quality control by comparison to modENCODE time course? eQTLs, was this sample used for the multi-stage eQTL study?

Bibliography

Abrams, John M, Kristin White, Liselotte I Fessler, and Hermann Steller (1993). "Programmed cell death during Drosophila embryogenesis." *Development (Cambridge, England)* 117.1, pp. 29–43.

Adams, Mark D, Susan E Celniker, Robert A Holt, Cheryl A Evans, Jeannine D Gocayne, Peter G Amanatides, Steven E Scherer, Peter W Li, Roger A Hoskins, et al. (2000). "The genome sequence of Drosophila melanogaster". *Science* 287.March, pp. 2185–95.

Albert, Frank W and Leonid Kruglyak (2015). "The role of regulatory variation in complex traits and disease". *Nature Reviews Genetics* 16.4, pp. 197–212.

Alexa, Adrian and Jorg Rahnenfuhrer (2010). *topGO: Enrichment analysis for Gene Ontology*. R package version 2.12.0.

Allison, David B, Xiangqin Cui, Grier P Page, and Mahyar Sabripour (2006). "Microarray data analysis: from disarray to consolidation and consensus." *Nature Reviews Genetics* 7.1, pp. 55–65.

Altenburg, Edgar and Hermann J Muller (1920). "The Genetic Basis of Truncate Wing,-an Inconstant and Modifiable Character in Drosophila." *Genetics* 5.1, pp. 1–59.

Altshuler, David, Mark J Daly, and Eric S Lander (2008). "Genetic Mapping in Human Disease". *Science* 322.5903, pp. 881–888.

Alwine, James C, David J Kemp, and George R Stark (1977). "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes." *Proceedings of the National Academy of Sciences of the United States of America* 74.12, pp. 5350–4.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber (2015). "HTSeq–a Python framework to work with high-throughput sequencing data". *Bioinformatics* 31.2, pp. 166–169.

Anderson, John S Jacobs and Roy Parker (1998). "The 3' to 5' degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SK12 DEVH box protein and 3' to 5' exonucleases of the exosome complex". *EMBO Journal* 17.5, pp. 1497–1506.

Anderson, Kathryn V, Gerd Jürgens, and Christiane Nüsslein-Volhard (1985). "Establishment of dorsal-ventral polarity in the Drosophila embryo: Genetic studies on the role of the Toll gene product". *Cell* 42.3, pp. 779–789.

Atwell, Susanna, Yu S Huang, Bjarni J Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M Tarone, et al. (2010). "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines." *Nature* 465.7298, pp. 627–631.

Ayroles, Julien F, Mary Anna Carbone, Eric a Stone, Katherine W Jordan, Richard F Lyman, Michael M Magwire, Stephanie M Rollmann, Laura H Duncan, Faye Lawrence, et al. (2009). "Systems genetics of complex traits in Drosophila melanogaster." *Nature genetics* 41.3, pp. 299–307.

Bailey, Timothy L, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble (2009). "MEME SUITE: tools for motif discovery and searching." *Nucleic acids research* 37.Web Server issue, W202–8.

Banerji, Julian, Sandro Rusconi, and Walter Schaffner (1981). "Expression of a β-globin gene is enhanced by remote SV40 DNA sequences". *Cell* 27.2, pp. 299–308.

Bannister, Andrew J and Tony Kouzarides (2011). "Regulation of chromatin by histone modifications." *Cell research* 21.3, pp. 381–395.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao (2007). "High-Resolution Profiling of Histone Methylations in the Human Genome". *Cell* 129.4, pp. 823–837.

Bartel, David P. (2009). "MicroRNAs: Target Recognition and Regulatory Functions". *Cell* 136.2, pp. 215–233.

Bassett, Andrew R, Charlotte Tibbit, Chris P Ponting, and Ji Long Liu (2013). "Highly Efficient Targeted Mutagenesis of Drosophila with the CRISPR/Cas9 System". *Cell Reports* 4.1, pp. 220–228.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289 –300.

Bennett, Brian J, Charles R Farber, Luz Orozco, Hyun Min Kang, Anatole Ghazalpour, Nathan Siemers, Michael Neubauer, Isaac Neuhaus, Roumyana Yordanova, et al. (2010). "A high-resolution association mapping panel for the dissection of complex traits in mice." *Genome research* 20.2, pp. 281–90.

Benoit, Beatrice, Chun Hua He, Fan Zhang, Sarah M Votruba, Wael Tadros, J Timothy Westwood, Craig a Smibert, Howard D Lipshitz, and William E Theurkauf (2009). "An essential role for the RNA-binding protein Smaug during the Drosophila maternal-to-zygotic transition." *Development (Cambridge, England)* 136.6, pp. 923–932.

Bogdan, Sven, Oliver Grewe, Mareike Strunk, Alexandra Mertens, and Christian Klämbt (2004). "Sra-1 interacts with Kette and Wasp and is required for neuronal and bristle development in Drosophila." *Development (Cambridge, England)* 131.16, pp. 3981–3989.

Botstein, David, Raymond L White, Mark Skolnick, and Ronald W Davis (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms." *American journal of human genetics* 32.3, pp. 314–31.

Boyle, Alan P, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford (2008). "High-resolution mapping and characterization of open chromatin across the genome." *Cell* 132.2, pp. 311–22.

Brem, Rachel B., Gael Yvert, Rebecca Clinton, and Leonid Kruglyak (2002). "Genetic Dissection of Transcriptional Regulation in Budding Yeast". *Science* 296.5568, pp. 752–755.

Brennecke, Julius, David R. Hipfner, Alexander Stark, Robert B. Russell, and Stephen M. Cohen (2003). "bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila". *Cell* 113.1, pp. 25–36.

Bridges, Calvin B (1916). "Non-Disjunction as Proof of the Chromosome Theory of Heredity". *Genetics* 1.1, pp. 1–52.

Bridges, Calvin B and Thomas H Morgan (1923). "The third-chromosome group of mutant characters of Drosophila melanogaster". *Carnegie Institute Publications* 327, p. 130.

Brody, Thomas (1999). "The Interactive Fly: gene networks, development and the Internet". *Trends in Genetics* 15.8, pp. 333–334.

Bulger, Michael and Mark Groudine (2011). "Functional and mechanistic diversity of distal transcription enhancers". *Cell* 144.3, pp. 327–339.

Bush, William S and Jason H Moore (2012). "Chapter 11: Genome-Wide Association Studies". *PLoS Computational Biology* 8.12.

Bushati, Natascha, Alexander Stark, Julius Brennecke, and Stephen M. Cohen (2008). "Temporal Reciprocity of miRNAs and Their Targets during the Maternal-to-Zygotic Transition in Drosophila". *Current Biology* 18.7, pp. 501–506.

Button, Katherine S, John P A Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma S J Robinson, and Marcus R Munafò (2013). "Power failure: why small sample size undermines the reliability of neuroscience." *Nature reviews neuroscience* 14.May, pp. 365–76.

Campos-Ortega, José A and Volker Hartenstein (1997). *The Embryonic Development of Drosophila melanogaster*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–405.

Cannavo, Enrico, Nils Kölling, Dermot Harnett, David Garfield, Francesco P Casale, Jacob F Degner, Hilary E Gustafson, Matt Davis, Oliver Stegle, et al. (2015). "Genetic and developmental regulation of expression levels and isoform diversity during embryogenesis". *Submitted*.

Carswell, Susan and James C Alwine (1989). "Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences." *Molecular and cellular biology* 9.10, pp. 4248–58.

Cartegni, Luca, Shern L Chew, and Adrian R Krainer (2002). "Listening to silence and understanding nonsense: exonic mutations that affect splicing." *Nature Reviews Genetics* 3.4, pp. 285–298.

Casanova, Jordi and Gary Struhl (1989). "Localized surface activity of torso, a receptor tyrosine kinase, specifies terminal body pattern in Drosophila." *Genes & development* 3.12B, pp. 2025–38.

Castle, W E, F W Carpenter, A H Clark, S O Mast, and W M Barrows (1906). "The Effects of Inbreeding, Cross-Breeding, and Selection upon the Fertility and Variability of Drosophila". *Proceedings of the American Academy of Arts and Sciences* 41.33, p. 731.

Cavaloc, Yvon, Cyril F Bourgeois, Liliane Kister, and James Stévenin (1999). "The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers." *RNA (New York, N.Y.)* 5.3, pp. 468–483.

Champely, Stephane (2015). *pwr: Basic Functions for Power Analysis*. R package version 1.1-2.

Chia, Ruth, Francesca Achilli, Michael F W Festing, and Elizabeth M C Fisher (2005). "The origins and uses of mouse outbred stocks." *Nature genetics* 37.11, pp. 1181–6.

Cho-Park, Park F and Hermann Steller (2013). "Proteasome regulation by ADP-ribosylation". *Cell* 153.3, pp. 614–627.

Churchill, G A and R W Doerge (1994). "Empirical threshold values for quantitative trait mapping". *Genetics* 138.3, pp. 963–971.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." *Fly* 6.2, pp. 80–92.

Colgan, Diana F and James L Manley (1997). "Mechanism and regulation of mRNA polyadenylation". *Genes & Development* 11.21, pp. 2755–2766.

Conaway, Ronald C and Joan Weliky Conaway (2011). "Function and regulation of the Mediator complex". *Current Opinion in Genetics and Development* 21.2, pp. 225–230.

Corder, E, A Saunders, W Strittmatter, D Schmechel, P Gaskell, G Small, A Roses, J Haines, and M Pericak-Vance (1993). "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families". *Science* 261.5123, pp. 921–923.

Csárdi, Gábor, Alexander Franks, David S Choi, Edoardo M Airoldi, and D Allan Drummond (2015). "Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast". *PLOS Genetics* 11.5. Ed. by Michael Snyder, e1005206.

Darwin, Charles (1859). *On the origin of species by means of natural selection.* London: John Murray.

Davis, Matthew P A, Stijn van Dongen, Cei Abreu-Goodger, Nenad Bartonicek, and Anton J Enright (2013). "Kraken: A set of tools for quality control and analysis of high-throughput sequence data". *Methods* 63.1, pp. 41–49.

Decker, Carolyn J and Roy Parker (1993). "A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation." *Genes & Development* 7.8, pp. 1632–1643.

Degner, Jacob F, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, et al. (2012). "DNase I sensitivity QTLs are a major determinant of human expression variation." *Nature* 482.7385, pp. 390–4.

Degner, Jacob F, John C Marioni, Athma A Pai, Joseph K Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K Pritchard (2009). "Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data". *Bioinformatics* 25.24, pp. 3207–3212.

Dembeck, Lauren M, Wen Huang, Michael M Magwire, Faye Lawrence, Richard F Lyman, and Trudy F C Mackay (2015). "Genetic Architecture of Abdominal Pigmentation in Drosophila melanogaster". *PLOS Genetics* 11.5. Ed. by Corbin D. Jones, e1005163.

Denny, Paul, Sally Swift, Frances Connor, and Alan Ashworth (1992). "An SRY-related gene expressed during spermatogenesis in the mouse encodes a sequence-specific DNA-binding protein." *The EMBO journal* 11.10, pp. 3705–3712.

Di Giammartino, Dafne Campigli, Kensei Nishida, and James L Manley (2011). "Mechanisms and consequences of alternative polyadenylation." *Molecular cell* 43.6, pp. 853–66.

Dimas, Antigone S, Samuel Deutsch, Barbara E Stranger, Stephen B Montgomery, Christelle Borel, Homa Attar-Cohen, Catherine Ingle, Claude Beazley, Maria Gutierrez Arcelus, et al. (2009). "Common regulatory variation impacts gene expression in a cell type-dependent manner." *Science (New York, N.Y.)* 325.5945, pp. 1246–1250.

Ding, Zhihao, Yunyun Ni, Sander W Timmer, Bum-kyu Lee, Anna Battenhouse, Sandra Louzada, Fengtang Yang, Ian Dunham, Gregory E Crawford, et al. (2014). "Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association". *PLoS Genetics* 10.11. Ed. by Greg Gibson, e1004798.

Dixon, Anna L, Liming Liang, Miriam F Moffatt, Wei Chen, Simon Heath, Kenny C C Wong, Jenny Taylor, Edward Burnett, Ivo Gut, et al. (2007). "A genome-wide association study of global gene expression." *Nature genetics* 39.10, pp. 1202–7.

Djuranovic, Sergej, Ali Nahvi, and Rachel Green (2012). "miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay." *Science (New York, N.Y.)* 336.6078, pp. 237–40.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras (2013). "STAR: Ultrafast universal RNA-seq aligner". *Bioinformatics* 29.1, pp. 15–21.

Dunn, Olive Jean (1959). "Estimation of the Medians for Dependent Variables". *The Annals of Mathematical Statistics* 30.1, pp. 192–197.

Easton, Douglas F, Karen A Pooley, Alison M Dunning, Paul D P Pharoah, Deborah Thompson, Dennis G Ballinger, Jeffery P Struewing, Jonathan Morrison, Helen Field, et al. (2007). "Genome-wide association study identifies novel breast cancer susceptibility loci." *Nature* 447.7148, pp. 1087–1093.

Edery, Isaac, Laurence J Zwiebel, Marie E Dembinska, and Michael Rosbash (1994). "Temporal phosphorylation of the Drosophila period protein." *Proceedings of the National Academy of Sciences of the United States of America* 91.6, pp. 2260–2264.

Edgar, Bruce A and Gerold Schubiger (1986). "Parameters controlling transcriptional activation during early drosophila development". *Cell* 44.6, pp. 871–877.

Ettwiller, Laurence, Benedict Paten, Mirana Ramialison, Ewan Birney, and Joachim Wittbrodt (2007). "Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation." *Nature methods* 4.7, pp. 563–565.

Faust, Joseph E, Avani Verma, Chengwei Peng, and James A Mcnew (2012). "An inventory of peroxisomal proteins and pathways in drosophila melanogaster". *Traffic* 13.10, pp. 1378–1392.

Felsenfeld, Gary and Mark Groudine (2003). "Controlling the double helix". *Nature* 421.6921, pp. 448–453.

Feuk, Lars, Andrew R Carson, and Stephen W Scherer (2006). "Structural variation in the human genome." *Nature Reviews Genetics* 7.2, pp. 85–97.

Fisher, Ronald Aylmer (1921). "Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk". *The Journal of Agricultural Science* 11.02, p. 107.

Fisher, Ronald Aylmer (1918). "The Correlation between Relatives on the Supposition of Mendelian Inheritance". *Philosophical Transactions of the Royal Society of Edinburgh* 52, pp. 399–433.

Fisher, Ronald Aylmer (1930). *The genetical theory of natural selection*. London: Oxford University Press.

Fisher, Ronald Aylmer and W A Mackenzie (1923). "Studies in crop variation. II. The manurial response of different potato varieties". *The Journal of Agricultural Science* 13.03, p. 311.

Foe, Victoria E and Bruce M Alberts (1983). "Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis." *Journal of cell science* 61, pp. 31–70.

Francesconi, Mirko and Ben Lehner (2014). "The effects of genetic variation on gene expression dynamics during development." *Nature* 505.7482, pp. 208–11.

Fuda, Nicholas J, M Behfar Ardehali, and John T Lis (2009). "Defining mechanisms that regulate RNA polymerase II transcription in vivo." *Nature* 461.7261, pp. 186–92.

Gaffney, Daniel J, Jean-Baptiste Veyrieras, Jacob F Degner, Roger Pique-Regi, Athma a Pai, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard (2012). "Dissecting the regulatory architecture of gene expression QTLs." *Genome biology* 13.1, R7.

Gallo, Steven M, Dave T Gerrard, David Miner, Michael Simich, Benjamin Des Soye, Casey M Bergman, and Marc S Halfon (2011). "REDfly v3.0: Toward a comprehensive database of transcriptional regulatory elements in Drosophila". *Nucleic Acids Research* 39.SUPPL. 1, pp. 1–6.

Galton, Francis (1909). *Memories of my life*. London, Methuen & co.

Gautier, Laurent, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry (2004). "affy–analysis of Affymetrix GeneChip data at the probe level." *Bioinformatics (Oxford, England)* 20.3, pp. 307–15.

Gawande, Bharat, Mark D Robida, Andrew Rahn, and Ravinder Singh (2006). "Drosophila Sex-lethal protein mediates polyadenylation switching in the female germline." *The EMBO journal* 25.6, pp. 1263–1272.

Gaziova, Ivana, Peter C Bonnette, Vincent C Henrich, and Marek Jindra (2004). "Cell-autonomous roles of the ecdysoneless gene in Drosophila development and oogenesis." *Development (Cambridge, England)* 131.11, pp. 2715–2725.

Gerrits, Alice, Yang Li, Bruno M. Tesson, Leonid V. Bystrykh, Ellen Weersing, Albertina Ausema, Bert Dontje, Xusheng Wang, Rainer Breitling, et al. (2009). "Expression quantitative trait loci are highly sensitive to cellular differentiation state". *PLoS Genetics* 5.10.

Ghavi-Helm, Yad, Felix A Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen E M Furlong (2014). "Enhancer loops appear stable during development and are associated with paused polymerase". *Nature* 512.7512, pp. 96–100.

Ghildiyal, Megha and Phillip D Zamore (2009). "Small silencing RNAs: an expanding universe." *Nature Reviews Genetics* 10.2, pp. 94–108.

Gibson, Ursula E, Christian A Heid, and P Mickey Williams (1996). "A novel method for real time quantitative RT-PCR." *Genome Research* 6.10, pp. 995–1001.

Gil, Anna and Nick J Proudfoot (1987). "Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation." *Cell* 49.3, pp. 399–406.

Goncalves, Angela, Sarah Leigh-Brown, David Thybert, Klara Stefflova, Ernest Turro, Paul Flicek, Alvis Brazma, Duncan T Odom, and John C Marioni (2012). "Extensive compensatory cis-trans regulation in the evolution of mouse gene expression". *Genome Research* 22.12, pp. 2376–2384.

Goode, Scott, Michael Melnick, Tze-Bin Chou, and Norbert Perrimon (1996). "The neurogenic genes egghead and brainiac define a novel signaling pathway essential for epithelial morphogenesis during Drosophila oogenesis." *Development (Cambridge, England)* 122.12, pp. 3863–79.

Graveley, Brenton R, Angela N Brooks, Joseph W Carlson, Michael O Duff, Jane M Landolin, Li Yang, Carlo G Artieri, Marijke J van Baren, Nathan Boley, et al. (2011). "The developmental transcriptome of Drosophila melanogaster." *Nature* 471.7339, pp. 473–9.

Gross, David S and William T Garrard (1988). "Nuclease hypersensitive sites in chromatin." *Annual review of biochemistry* 57, pp. 159–197.

Gudbjartsson, Daniel F, G Bragi Walters, Gudmar Thorleifsson, Hreinn Stefansson, Bjarni V Halldorsson, Pasha Zusmanovich, Patrick Sulem, Steinunn Thorlacius, Arnaldur Gylfason, et al. (2008). "Many sequence variants affecting diversity of adult human height." *Nature genetics* 40.5, pp. 609–15.

Gundelfinger, Eckart D and Norbert Hess (1992). "Nicotinic acetylcholine receptors of the central nervous system of Drosophila". *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1137.3, pp. 299–308.

Gusella, James F, Nancy S Wexler, P Michael Conneally, Susan L Naylor, Mary Anne Anderson, Rudolph E Tanzi, Paul C Watkins, Kathleen Ottina, Margaret R Wallace, et al. (1983). "A polymorphic DNA marker genetically linked to Huntington's disease". *Nature* 306.5940, pp. 234–238.

Halsey, Lewis G, Douglas Curran-Everett, Sarah L Vowler, and Gordon B Drummond (2015). "The fickle P value generates irreproducible results". *Nature Methods* 12.3, pp. 179–185.

Harbison, Susan T, Lenovia J McCoy, and Trudy F C Mackay (2013). "Genome-wide association study of sleep in Drosophila melanogaster." *BMC genomics* 14, p. 281.

Harding, Katherine, Timothy Hoey, Rahul Warrior, and Michael Levine (1989). "Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila." *The EMBO journal* 8.4, pp. 1205–12.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass (2010). "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities". *Molecular Cell* 38.4, pp. 576–589.

Hentges, Kathryn E and Monica J Justice (2004). "Checks and balancers: balancer chromosomes to facilitate genome annotation." *Trends in genetics : TIG* 20.6, pp. 252–9.

Herranz, Héctor, Ginés Morata, and Marco Milán (2006). "calderón encodes an organic cation transporter of the major facilitator superfamily required for cell growth and proliferation of Drosophila tissues." *Development (Cambridge, England)* 133.14, pp. 2617–2625.

Hsu, Patrick D, Eric S Lander, and Feng Zhang (2014). "Development and applications of CRISPR-Cas9 for genome engineering". *Cell* 157.6, pp. 1262–1278.

Huang, Guo-Jen Jen, Sagiv Shifman, William Valdar, Martina Johannesson, Binnaz Yalcin, Martin S Taylor, Jennifer M Taylor, Richard Mott, and Jonathan Flint (2009). "High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues." *Genome research* 19.6, pp. 1133–1140.

Huang, Wen, Andreas Massouras, Yutaka Inoue, Jason Peiffer, Miquel Ràmia, Aaron M Tarone, Lavanya Turlapati, Thomas Zichner, Dianhui Zhu, et al. (2014). "Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines." *Genome research* 24.7, pp. 1193–208.

Huang, Xuehui, Yan Zhao, Xinghua Wei, Canyang Li, Ahong Wang, Qiang Zhao, Wenjun Li, Yunli Guo, Liuwei Deng, et al. (2011). "Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm". *Nature Genetics* 44.1, pp. 32–39.

Innocenti, Federico, Gregory M Cooper, Ian B Stanaway, Eric R Gamazon, Joshua D Smith, Snezana Mirkov, Jacqueline Ramirez, Wanqing Liu, Yvonne S Lin, et al. (2011). "Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue." *PLoS genetics* 7.5, e1002078.

Ivanov, Dobril K, Valentina Escott-Price, Matthias Ziehm, Michael M Magwire, Trudy F C Mackay, Linda Partridge, and Janet M Thornton (2015). "Longevity GWAS Using the Drosophila Genetic Reference Panel". *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*.

Jan, Calvin H, Robin C Friedman, J Graham Ruby, and David P Bartel (2011). "Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs." *Nature* 469.7328, pp. 97–101.

Ji, Zhe, Ju Youn Lee, Zhenhua Pan, Bingjun Jiang, and Bin Tian (2009). "Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development." *Proceedings of the National Academy of Sciences of the United States of America* 106.17, pp. 7028–33.

Jonas, Stefanie and Elisa Izaurralde (2015). "Towards a molecular understanding of microRNA-mediated gene silencing". *Nature Reviews Genetics* 16.7, pp. 421–433.

Kalinka, Alex T, Karolina M Varga, Dave T Gerrard, Stephan Preibisch, David L Corcoran, Julia Jarrells, Uwe Ohler, Casey M Bergman, and Pavel Tomancak (2010). "Gene expression divergence recapitulates the developmental hourglass model." *Nature* 468.7325, pp. 811–814.

Kim-Ha, Jeongsil, Karen Kerr, and Paul M Macdonald (1995). "Translational regulation of oskar mRNA by Bruno, an ovarian RNA-binding protein, is essential". *Cell* 81.3, pp. 403–412.

Klein, Robert J, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, et al. (2005). "Complement factor H polymorphism in age-related macular degeneration." *Science (New York, N.Y.)* 308.5720, pp. 385–9.

Knezetic, Joseph A and Donal S Luse (1986). "The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro". *Cell* 45.1, pp. 95–104.

Kornblihtt, Alberto R, Ignacio E Schor, Mariano Alló, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J Muñoz (2013). "Alternative splicing: a pivotal step between eukaryotic transcription and translation." *Nature reviews molecular cell biology* 14.3, pp. 153–65.

Korte, Arthur, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg (2012). "A mixed-model approach for genome-

wide association studies of correlated traits in structured populations." *Nature genetics* 44.9, pp. 1066–71.

Lappalainen, Tuuli, Michael Sammeth, Marc R Friedländer, Peter A C 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, et al. (2013). "Transcriptome and genome sequencing uncovers functional variation in humans." *Nature* 501.7468, pp. 506–11.

Lawlor, Debbie A, Roger M Harbord, Jonathan A C Sterne, Nic Timpson, and George Davey Smith (2008). "Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology". *Statistics in Medicine* 27.8, pp. 1133–1163.

Leek, Jeffrey T, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael a Irizarry (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." *Nature Reviews Genetics* 11.10, pp. 733–739.

Leptin, Maria (1999). "Gastrulation in Drosophila: the logic and the cellular mechanisms". *The EMBO Journal* 18.12, pp. 3187–3192.

Lettre, Guillaume, Christoph Lange, and Joel N. Hirschhorn (2007). "Genetic model testing and statistical power in population-based association studies of quantitative traits". *Genetic Epidemiology* 31.4, pp. 358–362.

Levsky, Jeffrey M and Robert H Singer (2003). "Fluorescence in situ hybridization: past, present and future." *Journal of cell science* 116.Pt 14, pp. 2833–2838.

Li, Bo and Colin N Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." *BMC bioinformatics* 12, p. 323.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics (Oxford, England)* 25.14, pp. 1754–60.

Li, Yang, Olga Alda Alvarez, Evert W Gutteling, Marcel Tijsterman, Jingyuan Fu, Joost A G Riksen, Esther Hazendonk, Pjotr Prins, Ronald H A Plasterk,

et al. (2006). “Mapping determinants of gene expression plasticity by genetical genomics in C. elegans.” *PLoS genetics* 2.12, e222.

Link, Nichole, Paula Kurtz, Melissa O’Neal, Gianella Garcia-Hughes, and John M Abrams (2013). “A p53 enhancer region regulates target genes through chromatin conformations in cis and in trans”. *Genes & Development* 27.22, pp. 2433–2438.

Lippert, Christoph, Francesco Paolo Casale, Barbara Rakitsch, and Oliver Stegle (2014). “LIMIX: genetic analysis of multiple traits”. *bioRxiv*, pp. 0–26.

Listgarten, Jennifer, Carl Kadie, Eric E Schadt, and David Heckerman (2010). “Correction for hidden confounders in the genetic analysis of gene expression.” *Proceedings of the National Academy of Sciences of the United States of America* 107, pp. 16465–16470.

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, p. 550.

Luschnig, Stefan, Tilmann Bätz, Kristina Armbruster, and Mark a. Krasnow (2006). “serpentine and vermiform encode matrix proteins with chitin binding and deacetylation domains that limit tracheal tube length in Drosophila”. *Current Biology* 16.2, pp. 186–194.

Ma, Yue, Emily L Niemitz, Patricia A Nambu, Xiaoliang Shan, Charles Sackerson, Miki Fujioka, Tadaatsu Goto, and John R Nambu (1998). “Gene regulatory functions of Drosophila Fish-hook, a high mobility group domain Sox protein”. *Mechanisms of Development* 73.2, pp. 169–182.

MacArthur, Daniel G, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K Pickrell, et al. (2012). “A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes”. *Science* 335.6070, pp. 823–828.

Mackay, Trudy F C, Stephen Richards, Eric A Stone, Antonio Barbadilla, Julien F Ayroles, Dianhui Zhu, Sònia Casillas, Yi Han, Michael M Magwire, et al. (2012). “The Drosophila melanogaster Genetic Reference Panel.” *Nature* 482.7384, pp. 173–8.

Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, et al. (2015). "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". *Cell* 161.5, pp. 1202–1214.

Magnani, Luca, Jérôme Eeckhoute, and Mathieu Lupien (2011). "Pioneer factors: Directing transcriptional regulators within the chromatin environment". *Trends in Genetics* 27.11, pp. 465–474.

Magwire, Michael M, Daniel K Fabian, Hannah Schweyen, Chuan Cao, Ben Longdon, Florian Bayer, and Francis M Jiggins (2012). "Genome-wide association studies reveal a simple genetic basis of resistance to naturally coevolving viruses in Drosophila melanogaster." *PLoS genetics* 8.11, e1003057.

Maquat, Lynne E and Gordon G Carmichael (2001). "Quality control of mRNA function". *Cell* 104.2, pp. 173–176.

Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." *Genome research* 18.9, pp. 1509–17.

Massouras, Andreas, Sebastian M Waszak, Monica Albarca-Aguilera, Korneel Hens, Wiebke Holcombe, Julien F Ayroles, Emmanouil T Dermitzakis, Eric a Stone, Jeffrey D Jensen, et al. (2012). "Genomic variation and its impact on gene expression in Drosophila melanogaster." *PLoS genetics* 8.11, e1003055.

McGinnis, William, Richard L Garber, Johannes Wirz, Atsushi Kuroiwa, and Walter J Gehring (1984). "A homologous protein-coding sequence in drosophila homeotic genes and its conservation in other metazoans". *Cell* 37.2, pp. 403–408.

McManus, C Joel, Joseph D Coolon, Michael O Duff, Jodi Eipper-Mains, Brenton R Graveley, and Patricia J Wittkopp (2010). "Regulatory divergence in Drosophila revealed by mRNA-seq". *Genome Research* 20.6, pp. 816–825.

Mendel, Gregor (1866). "Versuche über Pflanzenhybriden". *Verhandlungen des Naturforschenden Vereines, Brünn* IV, pp. 3–47.

Montgomery, Stephen B, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis (2010). "Transcriptome genetics using second generation sequencing in a Caucasian population." *Nature* 464.7289, pp. 773–7.

Moore, Melissa J and Nick J Proudfoot (2009). "Pre-mRNA processing reaches back to transcription and ahead to translation." *Cell* 136.4, pp. 688–700.

Morgan, Thomas H (1911a). "An attempt to analyze the constitution of the chromosomes on the basis of sex-limited inheritance in Drosophila". *Journal of Experimental Zoology* 11.4, pp. 365–413.

Morgan, Thomas H (1911b). "Random segregation versus coupling in Mendelian inheritance". *Science (New York, N.Y.)* 34.873, p. 384.

Morgan, Thomas H (1910). "Sex limited inheritance in Drosophila". *Science (New York, N.Y.)* 32.812, pp. 120–122.

Morgan, Thomas H, Alfred H Sturtevant, Hermann J Muller, and Calvin B Bridges (1915). *The mechanism of Mendelian heredity*. New York: H. Holt and company, p. 288.

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, pp. 621–628.

Mullaney, Julienne M, Ryan E Mills, W Stephen Pittard, and Scott E Devine (2010). "Small insertions and deletions (INDELs) in human genomes". *Human Molecular Genetics* 19.R2, pp. 131–136.

Muller, Hermann J (1927). "Artificial transmutation of the gene". *Science (New York, N.Y.)* 66.1699, pp. 84–87.

Müller, Reto, Friedrich Altmann, Dapeng Zhou, and Thierry Hennet (2002). "The Drosophila melanogaster brainiac protein is a glycolipid-specific beta 1,3N-acetyl-glucosaminyltransferase." *The Journal of Biological Chemistry* 277.36, pp. 32417–20.

Murre, Cornelis, Patrick Schonleber McCaw, and David Baltimore (1989). "A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins". *Cell* 56.5, pp. 777–783.

Nam, Douglas Kyung, Sanggyu Lee, Guolin Zhou, Xiaohong Cao, Clarence Wang, Terry Clark, Jianjun Chen, Janet D Rowley, and San Ming Wang (2002). "Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription." *Proceedings of the National Academy of Sciences of the United States of America* 99.9, pp. 6152–6156.

Newton-Cheh, Christopher, Toby Johnson, Vesela Gateva, Martin D Tobin, Murielle Bochud, Lachlan Coin, Samer S Najjar, Jing Hua Zhao, Simon C Heath, et al. (2009). "Genome-wide association study identifies eight loci associated with blood pressure." *Nature genetics* 41.6, pp. 666–76.

Nica, Alexandra C, Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena Sekowska, Mary Travers, Simon Potter, Elin Grundberg, et al. (2011). "The architecture of gene regulatory variation across multiple human tissues: The muTHER study". *PLoS Genetics* 7.2, pp. 1–9.

Noyes, Marcus B, Xiangdong Meng, Atsuya Wakabayashi, Saurabh Sinha, Michael H Brodsky, and Scot A Wolfe (2008). "A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system". *Nucleic Acids Research* 36.8, pp. 2547–2560.

Nüsslein-Volhard, Christiane and Eric Wieschaus (1980). "Mutations affecting segment number and polarity in Drosophila". *Nature* 287.5785, pp. 795–801.

Ostrowski, Stephen, Herman A Dierick, and Amy Bejsovec (2002). "Genetic control of cuticle formation during embryonic development of Drosophila melanogaster". *Genetics* 161.1, pp. 171–182.

Padgett, Richard A, Paula J Grabowski, Maria M Konarska, Sharon Seiler, and Phillip A Sharp (1986). "Splicing of Messenger RNA Precursors". *Annual Review of Biochemistry* 55.1, pp. 1119–1150.

Painter, T. S. (1933). "A new method for the study of chromosome rearrangements and the plotting of chromosome maps". *Science* 78.2034, pp. 585–586.

Parker, Roy and Haiwei Song (2004). "The enzymes and control of eukaryotic mRNA turnover." *Nature structural & molecular biology* 11.2, pp. 121–127.

Pauli, Andrea, John L Rinn, and Alexander F Schier (2011). "Non-coding RNAs as regulators of embryogenesis." *Nature Reviews Genetics* 12.2, pp. 136–49.

Pearson, Karl (1901). "On lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, pp. 559–572.

Pearson, Karl (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". *Philosophical Magazine Series 5* 50.302, pp. 157–175.

Petretto, Enrico, Jonathan Mangion, Nicholas J Dickens, Stuart A Cook, Mande K Kumaran, Han Lu, Judith Fischer, Henrike Maatz, Vladimir Kren, et al. (2006). "Heritability and tissue specificity of expression quantitative trait loci." *PLoS genetics* 2.10, e172.

Phatnani, Hemali P and Arno L Greenleaf (2006). "Phosphorylation and functions of the RNA polymerase II CTD". *Genes & Development* 20.21, pp. 2922–2936.

Poulson, Donald F (1937). "Chromosomal Deficiencies and the Embryonic Development of Drosophila Melanogaster." *Proceedings of the National Academy of Sciences of the United States of America* 23.3, pp. 133–7.

Proudfoot, Nick J (2011). "Ending the message : poly (A) signals then and now". *Genes & development* 25, pp. 1770–1782.

Proudfoot, Nick J, Andre Furger, and Michael J Dye (2002). "Integrating mRNA processing with transcription". *Cell* 108.4, pp. 501–512.

Ptashne, Mark (2013). "Epigenetics: core misconcept." *Proceedings of the National Academy of Sciences of the United States of America* 110.18, pp. 7101–3.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ravi, Dashnamoorthy, Amy M Wiles, Selvaraj Bhavani, Jianhua Ruan, Philip Leder, and Alexander J R Bishop (2009). "A network of conserved damage survival pathways revealed by a genomic RNAi screen". *PLoS Genetics* 5.6.

Rifkin, Scott A, David Houle, Junhyong Kim, and Kevin P White (2005). "A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression." *Nature* 438.7065, pp. 220–223.

Rifkin, Scott A, Junhyong Kim, and Kevin P White (2003). "Evolution of gene expression in the Drosophila melanogaster subgroup." *Nature genetics* 33.2, pp. 138–44.

Ringrose, Leonie and Renato Paro (2004). "Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins." *Annual review of genetics* 38, pp. 413–443.

Ronderos, David S, Chun-Chieh Lin, Christopher J Potter, and Dean P Smith (2014). "Farnesol-detecting olfactory neurons in Drosophila." *J Neurosc* 34.11, pp. 3959–68.

Russell, Steven R H, Natalia Sanchez-Soriano, Charles R Wright, and Michael Ashburner (1996). "The Dichaete gene of Drosophila melanogaster encodes a SOX-domain protein required for embryonic segmentation." *Development (Cambridge, England)* 122.11, pp. 3669–3676.

Samuels, Mark E, Paul Schedl, and Thomas W Cline (1991). "The complex set of late transcripts from the Drosophila sex determination gene sex-lethal encodes multiple related polypeptides." *Molecular and cellular biology* 11.7, pp. 3584–3602.

Sax, Karl (1923). "The Association of Size Differences with Seed-Coat Pattern and Pigmentation in Phaseolus vulgaris." *Genetics* 8.6, pp. 552–560.

Schadt, Eric E, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, et al. (2003). "Genetics of gene expression surveyed in maize, mouse and man." *Nature* 422.6929, pp. 297–302.

Schena, Mark, Dari Shalon, Ronald W Davis, and Patrick O Brown (1995). "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray". *Science* 270.5235, pp. 467–470.

Shin, Chanseok and James L Manley (2004). "Cell signalling and the control of pre-mRNA splicing." *Nature reviews molecular cell biology* 5.9, pp. 727–738.

Skaer, Nick and Pat Simpson (2000). "Genetic Analysis of Bristle Loss in Hybrids between Drosophila melanogaster and D. simulans Provides Evidence for Divergence of cis-Regulatory Sequences in the achaete–scute Gene Complex". *Developmental Biology* 221.1, pp. 148–167.

Smale, Stephen T and James T Kadonaga (2003). "The RNA polymerase II core promoter." *Annual review of biochemistry* 72, pp. 449–479.

Small, Stephen, Adrienne Blair, and Michael Levine (1992). "Regulation of even-skipped stripe 2 in the Drosophila embryo." *The EMBO journal* 11.11, pp. 4047–57.

Smibert, Peter, Pedro Miura, Jakub O Westholm, Sol Shenker, Gemma May, Michael O Duff, Dayu Zhang, Brian D Eads, Joe Carlson, et al. (2012). "Global patterns of tissue-specific alternative polyadenylation in Drosophila." *Cell reports* 1.3, pp. 277–89.

Spitz, François and Eileen E M Furlong (2012). "Transcription factors: from enhancer binding to developmental control". *Nature Reviews Genetics* 13.9, pp. 613–626.

St Johnston, Daniel and Christiane Nüsslein-Volhard (1992). "The origin of pattern and polarity in the Drosophila embryo." *Cell* 68.2, pp. 201–219.

Stefani, Giovanni and Frank J Slack (2008). "Small non-coding RNAs in animal development." *Nature reviews molecular cell biology* 9.3, pp. 219–230.

Stegle, Oliver, Leopold Parts, Richard Durbin, and John Winn (2010). "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies." *PLoS computational biology* 6.5, e1000770.

Stegle, Oliver, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin (2012). "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses." *Nature protocols* 7.3, pp. 500–7.

Stranger, Barbara E, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, et al. (2007). "Population genomics of human gene expression." *Nature Genet* 39.10, pp. 1217–1224.

Struhl, Gary (1982). "Genes controlling segmental specification in the Drosophila thorax." *Proceedings of the National Academy of Sciences of the United States of America* 79.23, pp. 7380–4.

Sturtevant, Alfred H (1921). "A Case of Rearrangement of Genes in Drosophila." *Proceedings of the National Academy of Sciences of the United States of America* 7.8, pp. 235–7.

Sturtevant, Alfred H (1913). "The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association". *Journal of Experimental Zoology* 14.1, pp. 43–59.

Sturtevant, Alfred H (1959). *Thomas Hunt Morgan.* Washington D.C.: National Academy of Sciences, p. 293.

Stutz, Françoise and Elisa Izaurralde (2003). "The interplay of nuclear mRNP assembly, mRNA surveillance and export". *Trends in Cell Biology* 13.6, pp. 319–327.

Sul, Jae Hoon, Towfique Raj, Simone de Jong, Paul IW de Bakker, Soumya Raychaudhuri, Roel A Ophoff, Barbara E Stranger, Eleazar Eskin, and Buhm Han (2015). "Accurate and Fast Multiple-Testing Correction in eQTL Studies". *The American Journal of Human Genetics*, pp. 1–12.

Sutton, Walter S (1903). "The chromosomes in heredity". *Biological Bulletin* 4, pp. 231–251.

Takagaki, Yoshio, Rebecca L Seipelt, Martha L Peterson, and James L Manley (1996). "The polyadenylation factor CstF-64 regulates alternative process-

ing of IgM heavy chain pre-mRNA during B cell differentiation". *Cell* 87.5, pp. 941–952.

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, et al. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell." *Nature methods* 6.5, pp. 377–82.

The 1000 Genomes Project Consortium (2012). "An integrated map of genetic variation from 1,092 human genomes." *Nature* 491.7422, pp. 56–65.

The FlyBase Consortium (2014). "FlyBase 102 - Advanced approaches to interrogating FlyBase". *Nucleic Acids Research* 42.November 2013, pp. 780–788.

The Gene Ontology Consortium (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nature genetics* 25.1, pp. 25–29.

The GTEx Consortium (2013). "The Genotype-Tissue Expression (GTEx) project." *Nature genetics* 45.6, pp. 580–5.

The International HapMap Consortium (2005). "A haplotype map of the human genome." *Nature* 437.7063, pp. 1299–320.

The International HapMap Consortium (2007). "A second generation human haplotype map of over 3.1 million SNPs." *Nature* 449.7164, pp. 851–61.

The Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature* 447.7145, pp. 661–78.

Thisse, B, C Stoetzel, C Gorostiza-Thisse, and F Perrin-Schmitt (1988). "Sequence of the twist gene and nuclear localization of its protein in endomesodermal cells of early Drosophila embryos." *The EMBO journal* 7.7, pp. 2175–83.

Thomas, Sean, Xiao-Yong Li, Peter J Sabo, Richard Sandstrom, Robert E Thurman, Theresa K Canfield, Erika Giste, William Fisher, Ann Hammonds, et al. (2011). "Dynamic reprogramming of chromatin accessibility during Drosophila embryo development". *Genome Biology* 12.5, R43.

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* 7, pp. 562–78.

Tung, Jenny, Xiang Zhou, Susan C Alberts, Matthew Stephens, and Yoav Gilad (2015). "The genetic architecture of gene expression levels in wild baboons". *eLife* 4, pp. 1–22.

Turner, F Rudolf and A P Mahowald (1976). "Scanning electron microscopy of Drosophila embryogenesis". *Developmental Biology* 50.1, pp. 95–108.

Ulitsky, Igor, Alena Shkumatava, Calvin H Jan, Alexander O Subtelny, David Koppstein, George W Bell, Hazel Sive, and David P Bartel (2012). "Extensive alternative polyadenylation during zebrafish development." *Genome research* 22.10, pp. 2054–66.

Valdar, William, Leah C Solberg, Dominique Gauguier, Stephanie Burnett, Paul Klenerman, William O Cookson, Martin S Taylor, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint (2006). "Genome-wide genetic association of complex traits in heterogeneous stock mice." *Nature genetics* 38.8, pp. 879–87.

Valencia-Sanchez, Marco Antonio, Jidong Liu, Gregory J. Hannon, and Roy Parker (2006). "Control of translation and mRNA degradation by miRNAs and siRNAs". *Genes and Development* 20.5, pp. 515–524.

Veyrieras, Jean-Baptiste, Daniel J Gaffney, Joseph K Pickrell, Yoav Gilad, Matthew Stephens, and Jonathan K Pritchard (2012). "Exon-specific QTLs skew the inferred distribution of expression QTLs detected using gene expression array data." *PloS one* 7.2, e30629.

Veyrieras, Jean-Baptiste, Sridhar Kudaravalli, Su Yeon Kim, Emmanouil T Dermitzakis, Yoav Gilad, Matthew Stephens, and Jonathan K Pritchard (2008). "High-resolution mapping of expression-QTLs yields insight into human gene regulation." *PLoS genetics* 4.10, e1000214.

Vogel, Christine and Edward M Marcotte (2012). "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses". *Nature Reviews Genetics* 13.4, pp. 227–232.

Vonesch, Sibylle Chantal, David Lamparter, Trudy FC Mackay, Sven Bergmann, and Ernst Hafen (2015). "Genome-wide analysis reveals novel regulators of growth in Drosophila melanogaster". *bioRxiv*, pp. 1–15.

Wang, David G, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, et al. (1998). "Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome". *Science* 280.5366, pp. 1077–1082.

Wang, Shenqiu, Satish Arcot Jayaram, Johanna Hemphälä, Kirsten-André Senti, Vasilios Tsarouhas, Haining Jin, and Christos Samakovlis (2006). "Septate-junction-dependent luminal deposition of chitin deacetylases restricts tube elongation in the Drosophila trachea." *Current biology : CB* 16.2, pp. 180–5.

Weatherbee, Scott D, Georg Halder, Jaeseob Kim, Angela Hudson, and Sean Carroll (1998). "Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the Drosophila haltere." *Genes & development* 12.10, pp. 1474–82.

Weigmann, Katrin, Robert Klapper, Thomas Strasser, Christof Rickert, Gerd Technau, Herbert Jäckle, Wilfried Janning, and Christian Klämbt (2003). "FlyMove–a new way to look at development of Drosophila." *Trends in genetics : TIG* 19.6, pp. 310–1.

Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, et al. (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." *Nucleic acids research* 42.Database issue, pp. D1001–6.

Wilkening, Stefan, Vicent Pelechano, Aino I Järvelin, Manu M Tekkedil, Simon Anders, Vladimir Benes, and Lars M Steinmetz (2013). "An efficient method for genome-wide polyadenylation site mapping and RNA quantification." *Nucleic acids research* 41.5.

Wittkopp, Patricia J and Gizem Kalay (2011). "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence". *Nature Reviews Genetics* 13.1, pp. 59–69.

Wray, Gregory A (2007). "The evolutionary significance of cis-regulatory mutations." *Nature Reviews Genetics* 8.3, pp. 206–216.

Yalcin, Binnaz, Jérôme Nicod, Amarjit Bhomra, Stuart Davidson, James Cleak, Laurent Farinelli, Magne Oesteraas, Adam Whitley, Wei Yuan, et al. (2010). "Commercially available outbred mice for genome-wide association studies". *PLoS Genetics* 6.9.

Yin, Dingzi, Ping Huang, Jiarui Wu, and Haiyun Song (2014). "Drosophila protein phosphatase V regulates lipid homeostasis via the AMPK pathway". *Journal of Molecular Cell Biology* 6.1, pp. 100–102.

Yip, M L Richard, L Michele Lamka, and Howard D Lipshitz (1997). "Control of germ-band retraction in Drosophila by the zinc-finger protein HINDSIGHT." *Development (Cambridge, England)* 124.11, pp. 2129–2141.

Yoon, Oh Kyu and Rachel B Brem (2010). "Noncanonical transcript forms in yeast and their regulation during environmental stress." *RNA (New York, N.Y.)* 16.6, pp. 1256–67.

Yu, Jianming, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, et al. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." *Nature genetics* 38.2, pp. 203–8.

Zhang, Zhiwu, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, et al. (2010). "Mixed linear model approach adapted for genome-wide association studies." *Nature genetics* 42.4, pp. 355–60.

Zhou, Qiang, Tiandao Li, and David H Price (2012). "RNA polymerase II elongation control." *Annual review of biochemistry* 81, pp. 119–43.

Zhou, Xiang and Matthew Stephens (2014). “Efficient multivariate linear mixed model algorithms for genome-wide association studies.” *Nature methods* 11.4, pp. 407–9.