UNIVERSITY OF CALIFORNIA,
IRVINE

# Duplication and cis/trans regulatory variants: Evolutionary genomics perspectives on gene regulation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Xinwen Zhang

Dissertation Committee:
Assistant Professor J. J. Emerson, Chair
Professor Anthony Long
Associate Professor Ali Mortazavi

2019

# DEDICATION

To

my committee, my parents, and my friends

for their support and belief in my potential.

And to

Qiuxi

for his company in all these lonely years.

# TABLE OF CONTENTS

Page

**LIST OF FIGURES** **vi**

**LIST OF TABLES** **vii**

**ACKNOWLEDGMENTS** **viii**

**CURRICULUM VITAE** **ix**

**ABSTRACT OF THE DISSERTATION** **xi**

**1 Introduction** **1**
- 1.1 Phenotypic Variation and Genetic Variation . . . 1
- 1.2 Variation in Protein Sequence and Expression Pattern . . . 2
- 1.3 The Genetic Architecture of Gene Expression and *cis*/*trans* Mechanisms . . 2
- 1.4 Examples of Phenotypic Variation Resulting from Regulatory Variation . . . 4
- 1.5 The Allele-Specific-Expression Experiment . . . 4
- 1.6 The Main Problems with the Allele Specific Expression Experiment . . . 7
- 1.7 Benefits of Genome Assembly with Long Reads . . . 8
- 1.8 The Following Documents and My Projects . . . 8

**2 The Expression of the Incipient Duplicated Genes** **10**
- 2.1 Abstract . . . 10
- 2.2 Introduction . . . 11
- 2.3 Methods . . . 14
  - 2.3.1 *Drosophila melanogaster* strains and their genome assemblies . . . 14
  - 2.3.2 Identifying recent gene duplicates . . . 14
  - 2.3.3 Measuring the gene expression in two conditions with RNA-seq . . . 15
  - 2.3.4 Aggregate and independent measurement of paralog transcript abundance . . . 16
  - 2.3.5 Differential expression test on total expression level and paralog-specific expression level . . . 17
- 2.4 Results . . . 18
  - 2.4.1 Distribution of the 35 paralogs selected for further study . . . 18
  - 2.4.2 The functional distribution of the 35 genes . . . 19
  - 2.4.3 Pairs of paralog genes are usually differentially expressed . . . 19

2.4.4 Gene duplication usually leads to increased expression . . . . . 20
2.5 Discussion . . . . . 21
2.6 Acknowledgement . . . . . 22

**3 Inferring Genetic Architecture of Expression Variation 31**
3.1 Abstract . . . . . 31
3.2 Introduction . . . . . 32
3.3 Materials and Methods . . . . . 34
3.3.1 Yeast strains and preparation of hybrid samples . . . . . 34
3.3.2 DNA extraction and sequencing . . . . . 35
3.3.3 RNA extraction and sequencing . . . . . 35
3.3.4 Sequencing and assembly of YPS128 and RM11-1a genome . . . . . 36
3.3.5 Collect DNA/RNA read counts . . . . . 37
3.3.6 *Cis* variation estimation . . . . . 40
3.3.7 Generate null datasets lacking *cis*-variation . . . . . 46
3.3.8 Bootstrap *cis* variation estimation . . . . . 48
3.3.9 Data availability . . . . . 48
3.4 Results . . . . . 49
3.4.1 Assembly of reference genomes . . . . . 49
3.4.2 Allele-specific RNAseq . . . . . 49
3.4.3 The beta-binomial distribution models *cis*-expression sampling variation better than the binomial distribution . . . . . 50
3.4.4 *Cis* variation between YPS128 and RM11-1a strain is ubiquitous and often small in magnitude . . . . . 54
3.5 Discussion . . . . . 55
3.6 Acknowledgement . . . . . 57

**4 Inferring Compensatory Evolution of *cis* and *trans* Regulatory Variation 73**
4.1 Abstract . . . . . 73
4.2 Introduction . . . . . 74
4.3 Materials and Methods . . . . . 76
4.3.1 Synthesis hybrid and parental samples with no true *cis* variance . . . 76
4.3.2 Calculate *cis*/*trans* with both methods . . . . . 76
4.3.3 Calculate *cis*/*trans* correlation of both methods . . . . . 76
4.3.4 Correlation coefficient distribution of both methods . . . . . 77
4.3.5 Test on true data of hybrid and parental samples . . . . . 77
4.4 Results . . . . . 77
4.4.1 Statistical principles for the correlated error in two methods of measuring *cis*/*trans* effect . . . . . 77
4.4.2 The reduction of correlated error by the independent hybrid method . 80
4.5 Discussion . . . . . 81
4.6 Acknowledgment . . . . . 83

**5 Conclusion 87**

**Bibliography** **91**

# LIST OF FIGURES


Page

1.1 Architecture for gene expression . . . . . 3
1.2 *Cis* regulation of hind-limb in sticklebacks . . . . . 5
1.3 *Trans* regulation of pigmentation in *Petunia* . . . . . 5
1.4 Allele-specific experiment using RNA-seq . . . . . 6

2.1 Relative position of duplicated genes in three Strains . . . . . 24
2.2 The Go categories of the duplicated genes . . . . . 25
2.3 Paralog-specific read counts in two conditions . . . . . 26
2.4 FPKM of each paralog for 16 genes . . . . . 27
2.5 Total expression read counts in two conditions . . . . . 28
2.6 FPKM of total expression for 35 genes . . . . . 29
2.7 Distribution of log odds ratio . . . . . 30

3.1 Continuity of YPS128 and RM11-1a strain . . . . . 60
3.2 False positive rate with different number of replicates . . . . . 61
3.3 False positive rates of simulated data . . . . . 62
3.4 Discovery rate of the dataset "`Xinw2018`" . . . . . 63
3.5 Estimated confidence intervals of *cis* effect . . . . . 64
3.6 Distribution of *cis* effect . . . . . 65
3.7 Mapping bias of DNA counts . . . . . 67
3.8 Mapping of DNA counts with all SNPs . . . . . 68
3.9 Mapping of DNA counts without suspected SNPs . . . . . 69
3.10 Mapping of expression read counts . . . . . 70
3.11 Correlation of 40 expression profiles . . . . . 71
3.12 Statistical power of binomial model and beta-binomial model . . . . . 72

4.1 Correlation of *cis* and *trans* effect on simulated null dataset . . . . . 84
4.2 Correlation of *cis* and *trans* effect on experimental dataset . . . . . 85
4.3 Work flow for the synthesis of hybrid and parental samples . . . . . 86

# LIST OF TABLES

Page

3.1 Summary of the datasets . . . . . 58
3.2 Contiguity, completeness, and accuracy of YPS128 and RM11-1a genomes . . 59
3.3 Number of genes and their proportion with different *cis*-effect magnitude . . 66

# ACKNOWLEDGMENTS

I want to thank my parents for supporting my decision to go aboard and do whatever I think it's deserved to do. Their persistence in self-improving, respect to knowledge, rationality in that irrational social environment affect me for many years. I am so fortunate to be their child.

A special thanks to my very first academic advisor in China Agriculture University, professor Huiqiang Lou, for his guidance and encouragement. He welcomed me to his molecular biology lab even though I didn't know how to use a pipette. Most importantly, he directed my interests to evolution. Without him, I don't think I could stay in academics.

I also want to thank the "Mathematical, computational and systems biology" program, who accepted me and made all these complicated "go abroad" things possible. The sympathetic MCSB staffs created an amiable atmosphere that I didn't feel any culture shock. I appreciate the quantitative training the program offers and all the intellectual inputs from my smart classmates of diverse backgrounds.

My most sincere and deepest gratitude goes to my thesis advisor, J. J. Emerson, for his guidance throughout these years. Although I admire him for his academic achievement, talent, and abundant knowledge across several different subjects, he is more like a friend other than a role model high above. He is willing to share his good or bad experiences in science, teaching, and grant writing. Those experiences not only entertain me but also comfort me by knowing that life is hard even for professors but still fun. It's very fortunate for me to be his first graduate student and witness the gradual growth of the lab. I gained many facets of abilities during the process, both scientific and non-scientific.

I would also like to thank my friends and colleagues who assist me in these projects. Mahul Chakraborty, Roy Zhao, Luna Ngo, Yi Liao, Arun Ramaiah, James Baldwin-Brown, Tatsuhisa Tsuboi, Mandy Jiang and many others contributed in various ways.

My sincere thanks go to Dr. Ali Mortazavi and Dr. Anthony Long, for being my committee, listening to my ideas, and providing their perspectives on my projects. The projects are made possible by the fellowships and TAships from the School of Biological Science, UC Irvine Startup Fund and NIH grant R01GM123303-1 awarded to J.J. Emerson. I own thanks to the staff in EcoEvo department for all they did to make the process smooth.

Part of Chapter 2 is published in *Trends in Genetics*. Chapter 3 is a preprint in *bioRxiv*. Part of Chapter 4 is published in *Nature Genetics*. I want to thank those journals for allowing the reprint of the articles in this dissertation.

# CURRICULUM VITAE

## Xinwen Zhang

### EDUCATION

**Doctor of Philosophy in Biological Sciences** **2019**
University of California, Irvine *Irvine, California*

**Mathematical, Computational and Systems Biology Program** **2014**
University of California, Irvine *Irvine, California*

**Bachelor of Science in Biological Sciences** **2013**
China Agricultural University *Beijing, China*

### RESEARCH EXPERIENCE

**Graduate Research Assistant** **2013–2019**
University of California, Irvine *Irvine, California*

**Undergraduate Research Assistant** **2012–2013**
China Agricultural University *Beijing, China*

### TEACHING EXPERIENCE

**Teaching Assistant** **2015–2018**
Department of Ecology and Evolutionary Biology, UC Irvine *Irvine, California*

**BIO94** - Genetic Analysis (2 quarters)
**BIO97** - Organisms to Ecosystems (3 quarters)
**E115L** - Evolution Lab (4 quarters)

**Undergraduate Research Mentor** **2016–2018**
Undergraduate Research Opportunities Program, UC Irvine *Irvine, California*

## REFEREED JOURNAL PUBLICATIONS

**X. Zhang**, J. J. Emerson. Inferring Compensatory Evolution of cis- and trans- Regulatory Variation. Trends Genet 35(1):1–3. 2019

**X. Zhang**, J. J. Emerson. Inferring the genetic architecture of expression variation from replicated high throughput allele-specific expression experiments. bioRxiv:699074. 2019

M. Chakraborty, N. W. VanKuren, R. Zhao, **X. Zhang**, S. Kalsow, and J. J. Emerson. Hidden genetic variation shapes the structure of functional elements in Drosophila. Nat. Genet. 50, 20–25. 2018

## REFEREED CONFERENCE PUBLICATIONS

Lightning Talk: Inferring the genetic architecture of expression variation from replicated high throughput allele-specific expression experiments Aug 2018

The Yeast Genetics Meeting, Palo Alto, CA

Talk: The evolution of transcriptional regulation Jan 2018

Grad Presentation for Recruitment, UC Irvine

Talk: Statistical inference on allele specific expression Feb 2016

Winter EEB Graduate Students Symposium, UC Irvine

Poster: Statistical inference on allele specific expression July 2016

The Allied Genetics Conference, Orlando, FL

Poster: Statistical inference on allele specific expression Feb 2015

Southern California Evolutionary Genetics Meeting, UC Riverside

# ABSTRACT OF THE DISSERTATION

Duplication and cis/trans regulatory variants: Evolutionary genomics perspectives on gene regulation

By

Xinwen Zhang

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2019

Assistant Professor J. J. Emerson, Chair

Variation in gene expression contributes significantly to phenotypic variation. As a result, in addition to protein-coding loci, and genomic regions coding for gene regulatory elements are predicted to be under selection. This dissertation uses the genetic models budding yeast and fruit fly to explore the genetic basis of gene expression variation within species. The first chapter lays out the general background. It introduces the genetic architecture of gene expression, the *cis/trans* model, the allele-specific expression approach, and other commonly used methods in genomic studies.

The second chapter explores the relation of gene duplication and gene expression level. Gene duplication is thought to be the primary mechanism to produce new genes. However, a newly duplicated gene copy needs to exist in the population long enough to gain novel function. If the new copy affects gene expression in a deleterious direction, it would soon be eliminated by purifying selection. We would like to know how gene duplication affects expression between the duplicated genotype and the single copy genotype as well as the differences between paralogs in the duplicated genotype. We compared the genomes of strains of *Drosophila melanogaster*, focusing on 35 newly duplicated nuclear genes and compared the gene expression level between two duplicated paralogs and between the singletons and

doublets. We found that all of the 16 analyzable genes show differential expression between paralogs under the binomial model. The other 19 genes are either 100% identical in their sequence or have more than two duplicated copies, rendering analysis of copy-specific expression patterns either impossible or ambiguous. For the total expression level, we found that most of the genes show elevated expression level, though the magnitude of change shows no clear relationship with the number of copies. The work is the first such genome-wide survey of duplicated gene expression employing comparisons of reference-grade genome assemblies. This ensures that we discover duplicates previously hidden to short-read based methods.

The third chapter discusses a novel implementation of statistical model for inference of allele-specific expression. Commonly used binomial models ignore the variance among biological replicates which leads to many false-positives. We implemented a beta-binomial model and demonstrated its advantages with both simulated and experimental data. The 20 biological replicate allele-specific expression dataset not only yields a more accurate landscape of expression variation but also provide a resource for model testing for future studies.

The fourth chapter contributes to a debate regarding the commonly reported compensatory evolution in expression regulatory control. We demonstrate with statistical principles that the observed compensatory evolution in allele specific expression studies might merely be a measurement artifact. It then discusses an improved method and demonstrates the reduction of the negative-correlation (an indicator of compensatory evolution) mediated by shared error. Inferences are made with both simulated and published data.

The fifth and final chapter is a summary of the thesis. It also points out several unsolved problems and put forward future directions.

# Chapter 1

# Introduction

This chapter outlines a general background for the dissertation. It introduces the genetic architecture of gene expression, the *cis/trans* model, the allele-specific experiment and interpretation, and other commonly used methods in genomic studies.

## 1.1 Phenotypic Variation and Genetic Variation

The analysis of how the phenotypic variation corresponds to the genetic variation is a long-standing task in genetics. How genotype relates to phenotype can be studied in both directions. One direction is from the genotype: certain parts of the genetic material (DNA) are manipulated, and consequences in phenotype are then recorded. The other direction starts with phenotypes. We start from the natural variation of phenotype and then by comparing the underlying DNA sequences, map the phenotype to DNA locations. Those natural variations in phenotypes, unlike the manipulated ones, are more likely to be evolution heritage. For evolutionists, natural variation is of primary interest.

## 1.2 Variation in Protein Sequence and Expression Pattern

The central dogma claims that information from DNA passes to RNA via transcription and then into protein via translation (and once in protein, cannot go back). The protein is generally thought to constitute the major driver of phenotype, indicating that both gene sequence and gene expression variance are relevant for variation in phenotype.

The opinion that variation in gene expression contributes significantly to phenotypic variation has been proposed by many early studies [22, 42]. One widely cited study compares human and Chimpanzee proteins, pointing out that the peptide sequences are almost the same. This unexpectedly small variation in protein sequence was argued to be insufficient to explain the phenotypic differences between the species [26]. Consequently, the authors proposed that variation in gene expression as a candidate explaining the phenotypic difference. Consequently, gene regulatory elements have long been thought to be an important target of natural selection comparable in significance to variation in the proteome [47, 76].

## 1.3 The Genetic Architecture of Gene Expression and *cis*/*trans* Mechanisms

The expression of genes can be seen as a signal transmission process. The molecules in the cellular environment occupied by the chromosomes (e.g., transcription factors) provide a signal input and the sequence features linked to the gene itself detect the signal (Figure 1.1). A typical eukaryotic gene has several regulatory regions upstream of the transcription starting point. The proximal regulatory sequence, also known as a promoter, is where the RNA polymerase bind to initiate the transcription process. The other distal regulation

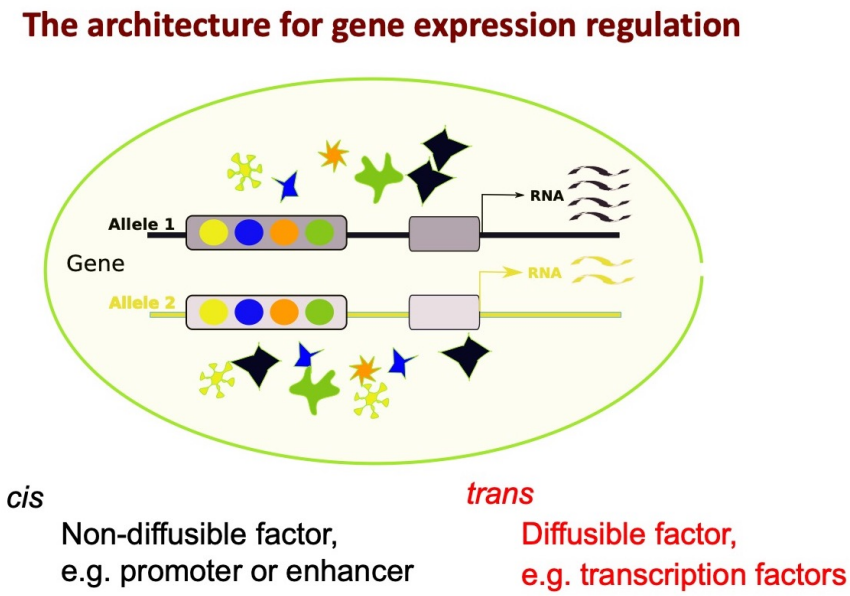



Figure 1.1: The *cis* and *trans* factors affecting gene expression are shown as colorful elements on the sequence and colorful surrounding elements. The promoter region is shown as a gray rectangle upstream of the transcription starting point.

sequences are for complex time and space control. These regions are named enhancer or insulator depending on their function. They detect signals (e.g., transcription factors with cell cycle information or spatial information) and guide the timing and amount of expression by attracting or repelling RNA polymerase.

If the variation in DNA sequence affects the expression level in a non-diffusible manner, we call it *cis* variation. For example, a mutation in the promoter region might reduce RNA polymerase binding leading to reduced expression. Another example: a deletion in the enhancer region that makes the gene insensitive to some growth factor. These are DNA sequence variations, and would not diffuse out to affect other chromosomes. On the other hand, if the variation at the DNA sequence level affects the expression level in a diffusible manner, we call it *trans* variation – for example, an amino acid change in a transcription factor. The transcription factor is diffusible and has the chance to affect genes in different chromosomes. We can decompose the genetic architecture of regulatory variation into *cis* variation and *trans* variation. Their magnitude can be measured by the allele-specific expression experiment

introduced in section 1.5.

## 1.4 Examples of Phenotypic Variation Resulting from Regulatory Variation

Mutations that affect gene expression in a non-diffusible manner are called *cis* variants. Such mutations include nucleotide changes to promoters and enhancers as well as changes in copy number of a gene. Additionally, mutations like TE insertions near the gene that change the chromatin state and inversions that change the upstream and downstream context of a gene can affect the expression of genes in a *cis* manner. The *trans* variation includes even broader events; any mutation that changes the micro-environment in which the gene exists is potentially effective.

Gene expression variation sometimes results in variation in observable phenotype. A classical phenotype of the hind-limb reduction in three-spined sticklebacks is caused by *cis* mutation [60]. A regulatory mutation on the upstream of the Pitx1 gene reduces or eliminates the Pitx1 gene expression in pelvic and caudal fin precursors (Figure 1.2).

The pigmentation of *Petunia* is an example of *trans* variant with phenotypic consequences. The expression level of several pigmentation genes are affected by a frameshift mutation of an2 transcription factor (Figure 1.3).

## 1.5 The Allele-Specific-Expression Experiment

The allele-specific-expression experiment is similar to a standard gene expression experiment. We first extract mRNA, make a library, sequence it, then map the reads to genes

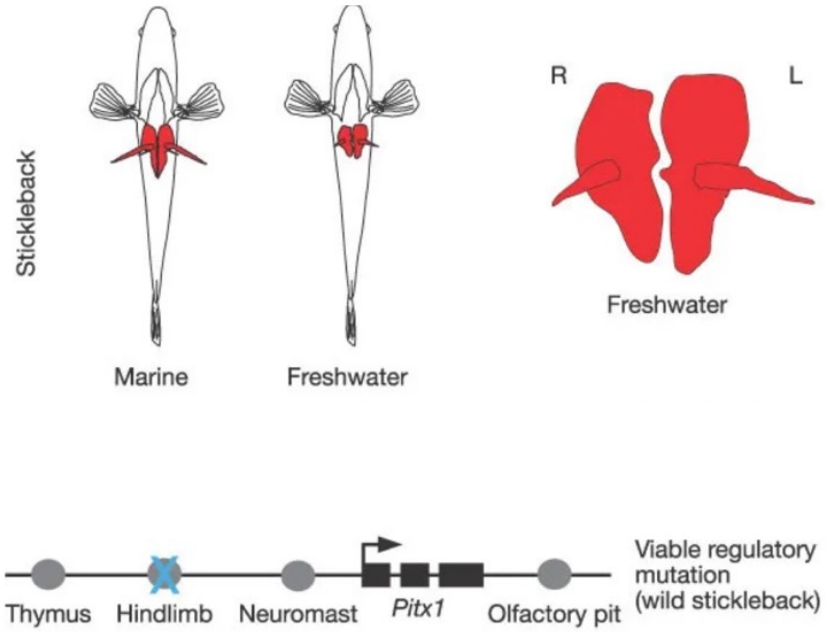



Figure 1.2: A mutation happens on the hind-limb regulatory region upstream of the Pitx1 gene eliminates the expression of Pitx1 and leads to hind-limb reduction in freshwater sticklebacks.

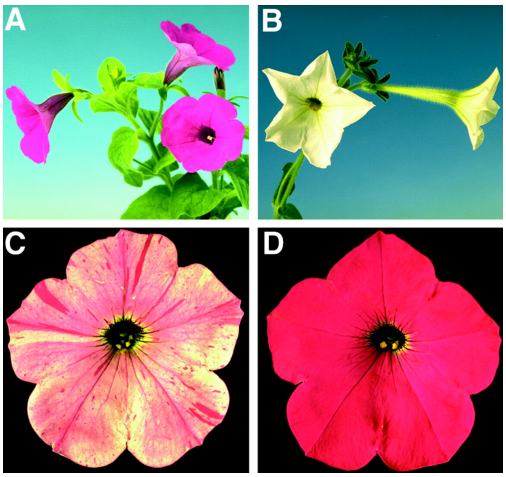



Figure 1.3: The Petunia have different flower color. The frameshift mutation on the transcription factor an2 change the expression of several genes involving in the pigmentation pathway.

in the genome, and finally calculate the read counts. Allele-specific-expression requires one additional step to identify from which allele an RNA read comes (for diploid hybrid cells or mixed homologous parental samples). Only reads overlapping the variant positions (SNPs or indels) between two alleles can be identified (Figure 1.4). Other reads are uninforma-

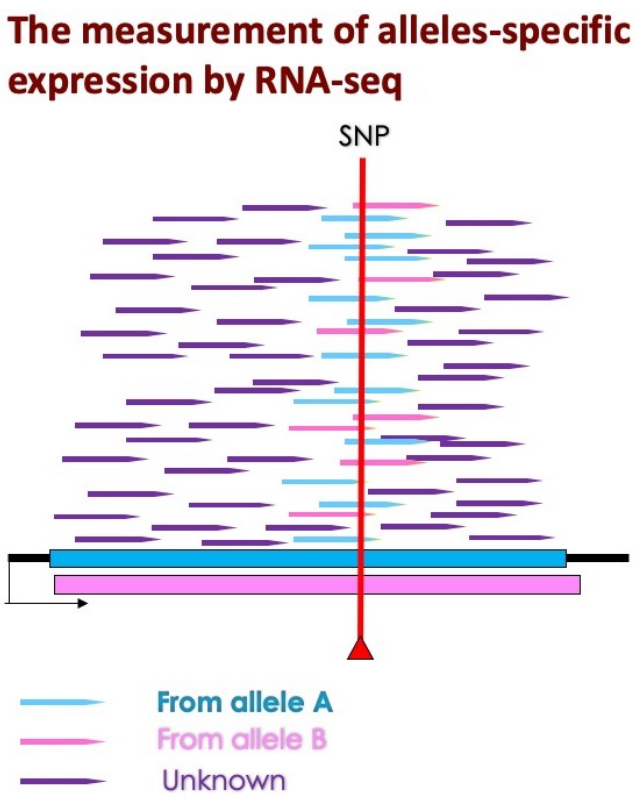



Figure 1.4: In allele-specific experiment, only reads overlapping with the variant positions (red line in figure) are informative of their parental origin. Other reads have to be discarded.

tive, so allele-specific expression experiments make use of a smaller proportion of collected data relative to standard gene expression experiments, and are therefore less powerful for differential expression inference given the same sequencing depth.

However, if the allele-specific read counts can be accurately collected, we can then calculate the *cis* and *trans* variation. In an F1 hybrid cell, two alleles of the same gene are exposed to the same diffusible elements, so any difference between the alleles' expression must be encoded by features linked to the gene itself (i.e., *cis* variation). By measuring the allele-specific expression of all genes in hybrid cells, we can measure the magnitude of *cis* variation (*cis*-effect). On the other hand, in an equal-mixed sample of parental cells, the allelic expression difference is caused by a combination of *cis* and *trans* effect. The *trans* effect can then be estimated accordingly [62, 15].

## 1.6 The Main Problems with the Allele Specific Expression Experiment

Although the allele-specific expression experiment is useful for characterizing *cis* and *trans* variation, it has some limitations. The major one is the mapping bias problem. As discussed in the previous section, only read counts overlapping with variant positions are informative of which alleles they come from. When only one reference genome is provided in the mapping step, the parental genome that is more similar to the reference genome tends to get more read counts. As a result, the mapping step is affected by the choice of the reference genome.

Without the ability to sequence and assemble the whole genomes, researchers tend to use the community reference genomes from public databases and mask the known variant positions or substitute SNPs for the nucleotides of the reference. It helps to reduce the mapping bias, but the variant list may not reflect the accurate variant positions for the specific strains in each experiment. Fortunately, with new sequencing and assembly techniques, high-quality genomes become accessible for individual labs. The problem can then be solved by using the experimental-specific reference genomes.

Another problem is the over-dispersed variance among biological replicates. Gene expression itself is a noisy process, not to mention the variance introduced in sample collection, library preparation, and sequencing. Previous studies use only 1-3 replicates and a binomial model, leading to many false positives of differential expression between alleles. Fortunately, as the library preparation and sequencing become more and more affordable, collecting more replicates becomes more feasible. A new scheme is described in Chapter 3.

## 1.7 Benefits of Genome Assembly with Long Reads

The newest mainstream generation of sequencing platforms like Single Molecule Real Time Sequencing by Pacific Biosciences and the MinION Nanopore sequencer by Oxford Nanopore Tech make high-quality genomes accessible for individual labs. My lab colleagues developed a hybrid assembly scheme that combines the high contiguity from long reads and the high accuracy from short reads while keeping sequencing cost as low as possible [9]. With the help of these new techniques and schemes, the genomes of the two *Saccharomyces cerevisiae* strain: RM11-1a and YPS128 and the genomes of two *Drosophila melanogaster* strain: A3 and A4 are accurately sequenced and assembled.

The RM11-1a and YPS128 genomes are both used as the reference genome in the allele-specific-expression experiments so that mapping bias is eliminated (For details, see chapter 3).

A3 and A4 genomes are compared with the ISO1 *Drosophila melanogaster* release 6 reference to find duplicated genes in Chapter 2. Duplicates are very hard to identify with only short-read assemblies. This project would be tedious and filled with uncertainty from inaccurate assemblies without these techniques. Reference quality genomes will undoubtedly become routine for genomic studies going forward. The struggle to deal with the uncertainty in genomes could soon be saved by such techniques.

## 1.8 The Following Documents and My Projects

This dissertation focuses on the natural variation in gene regulation mechanisms and attempts to solve the main problems in commonly used methods by taking advantage of state-of-the-art genome sequencing and assembly methods. The new methods/schemes are then

tested on simulated data and also applied to real experimental data, yielding a preliminary landscape of regulatory variation.

The first project focuses on the duplicated genes in two strains of *Drosophila melanogaster*. The copy number differences of genes between these two strains can be viewed as one type of *cis* variation. The expression levels of genes with different copy numbers are compared within and between the two strains. This work is described in Chapter 2. As of this writing, it is unpublished, but I anticipate a manuscript based on this work will be submitted soon after submission of this dissertation. Some of of the results related to this chapter are already published in *Nature Genetics*.

For general *cis* measurements, the statistical methods should use a model that incorporates over-dispersion among replicates. In the second project, I outline a beta-binomial model for general *cis* variation measurement and applied to 20 replicated yeast hybrid samples. Along the way, I also solved the mapping bias problem with a carefully-designed pipeline involving *de novo* assembly and identification and correction of biased variants. This work is preprinted in bioRxiv and is described in Chapter 3.

Compensatory evolution in *cis* and *trans* gene regulation has been reported in several studies. However, many studies employ a method for *cis*/*trans* estimation that conflates compensatory evolution with shared error. The third project demonstrates the phenomena with both statistical prove and simulated/experimental data. Part of this work is published in Trends in Genetics and is fully described in Chapter 4.

# Chapter 2

# The Expression of the Incipient Duplicated Genes

This project explores the relationship between gene duplication and gene expression level. This comprehensive genome-wide survey of duplicated gene expression is made possible with newly-developed genome sequencing and assembly approaches. Part of the work is published in *Nature Genetics* under the title "Hidden genetic variation shapes the structure of functional elements in Drosophila". This chapter includes additional follow-up work (including more extensive analysis that addresses all duplicates rather than a few examples) slated for submission as an independent manuscript.

## 2.1 Abstract

Gene duplication plays an important part in genome evolution. Here we measure the gene expression of 35 newly duplicated genes in two strains of *Drosophila melanogaster*: A3 and A4. Each of these genes is fully duplicated. Of the 35 copied genes, 16 exhibit allelic dif-

ferences making them suitable for paralog-specific expression analysis. We found that the paralogs are usually differentially expressed, though the change is often less than two-fold. Provided both copies are expressed, the combined expression levels of the two copy genotypes tend to increase relative to that of the corresponding single copy genotypes. However, the magnitude of change has no clear pattern. While in a few cases, the total expression level corresponds roughly linearly to gene copy (ie doubling for a one-copy to two-copy comparison), there are many exceptions. Indeed, these cases offer many examples where at least one of the apparently redundant individual copies does not faithfully recapitulate the expression of its single copy progenitor. Each copy number variant on this list merits careful investigation of its regulatory mechanism to understand how these changes in expression came about mechanistically.

## 2.2 Introduction

Gene duplication is one of the most important mechanisms for adding novel genes to the genome. Members of gene families can often easily be traced back to a gene ancestor [46]. Duplication could happen via ectopic recombination, replication slippage, errors in double strand break repair, retrotransposition, or whole genome duplication, the consequences of which may occasionally be adaptive [27] or disease-causing [64]. One apparent dilemma for the fixation of the new duplication is attaining novel function that would preserve the new copy may be slow compared to loss by mutation, genetic drift, or purifying selection. While there are many models proposed to explain how duplicates are fixed and retained in populations [21], most imply that the fixed duplicates are neutral or beneficial.

We identified newly duplicated genes in both A3 and A4 genomes. We restricted our attention to loci for which the reference genome ISO1 possessed a single copy allele. Since these copy number changes are present in either A3 or A4, the duplication event will usually have

occurred since A3 and A4 last shared a common ancestor at the locus. But since variation at the locus is still segregating, these duplicates' final fate remains unresolved. Any effects of gene duplication on the expression level potentially affect the fitness effects of the duplicate allele. It is possible that for one gene, a dosage increase is beneficial but for another gene, an increase is deleterious. Environment is also a factor. An increase of expression could be beneficial in an environment the A3 strain most recently inhabited, but could be deleterious in the corresponding A4 environment.

Since the connection between expression level and fitness is *a priori*, predicting the fitness effect and the likely fate of duplication is not tractable given our approach. Instead, we focus on surveying the molecular phenotypes of expression level change arising from gene duplications. Since duplicate discovery is conditional on a small sample size, any ascertained variants are drawn from intermediate to high allele frequency classes. As a result, they are disproportionately likely to be beneficial relative to variants ascertained in the absence of this bias. Interestingly, some genes also exhibit a clear response to a nicotine treatment and fall under nicotine resistance QTL, which are likely to be positively selected in the area where nicotine (e.g. tobacco) or neo-nicotinoid pesiticides are prevalent [41, 11].

Previous results appear contradictory about effect of duplication on expression level. Gene duplicates are variously reported to increase expression levels [37], decrease noise in expression levels [55], or maintain consistent expression levels [56]. We hope this work will serve as a rigorous quantitative estimate of the distribution of these possible outcomes, and that our measurement of paralog specific expression will provide mechanistic insight into various outcomes.

In this project, we identified 35 genes that have no recent paralogs in the ISO1 strain, but do exhibit recent paralogs in either the A3 or the A4 genomes. These genes have at least 95% identity in their protein coding regions, so are presumably recent duplicates. They are distributed across all 5 chromosome arms, and are comprised mostly of tandem

duplications, with only 2 translocation duplicates. A GO analysis shows that these genes are over-represented by the "UDP-glycosyltransferase activity" class. Out of the 35 genes, 16 exhibit sufficient nucleotide variation between the copies to subject them to paralog-specific expression analysis. Naive statistical tests of differential expression between paralogs show that all 16 pairs are differentially expressed between the two paralogs with a wide variance in the degree of change. An important caveat is that, since we only have one biological replicate for this experiment, the number of differentially expressed pairs would be overestimated by this simple binomial approach. Our expression dataset from Marriage et al. 2014 (a nicotine resistance QTL study) measured expression in two conditions: larvae with and without nicotine in the medium. Most of the pairs did not change expression pattern across conditions (the most highly expressed paralog in the control condition will usually be more highly expressed in the nicotine condition as well), indicating the variance in the *cis* regulatory region or the position effect is consistently behaved or independent of the *trans* regulatory network. We analyzed all genes for differences in the total expression level between singleton and multi-copy alleles. The expression levels scale with copy number for some genes, but most exhibit changes that deviate from this ideal, showing either higher or lower than the expected variation. This demonstrates that the total expression level is not merely driven by doubling the genetic material, and must entail some regulatory network control. Future work carefully characterizing all of these variants will substantially improve our understanding of the spectrum of possibilities available for the molecular phenotype of gene expression following gene duplication and the regulatory mechanisms that underlie such changes.

## 2.3 Methods

### 2.3.1 *Drosophila melanogaster* strains and their genome assemblies

A3 and A4 strains are founder strains for the *Drosophila* Synthetic Population Resource (DSPR) project [25]. Regardless of their origins, the strains of the resource are cosmopolitan. A3 is from Spain whereas A4 is from Zimbabwe. The genomes of A3 and A4 have been assembled *do novo* from high-coverage PacBio data, resulting in reference quality genomes. The sequencing and assembly of these strains is described here [10].

### 2.3.2 Identifying recent gene duplicates

To simplify the problem, we only choose genes that meet the following criteria: 1) the gene is present as a single copy in the ISO1 reference genome and two or more copies in either the A3 or A4 genome; 2) the copy number differs between A3 and A4; 3) the protein-coding sequences (CDS) of a gene should all be present in all copies, meaning that partial duplication is not considered; 4) the copies should be similar to each other so that we only consider likely recent duplicates (compared to ISO1, CDSs exhibit $\geq$ 95% identity, retain $\geq$ 95% of CDS length, and retain $\geq$ 50% of the full gene span, which permits some variation in UTRs and intron sequences).

To achieve this, we first extract all CDS sequences from the ISO1 reference genome and its annotation file. We then blasted [7] the ISO1 CDS sequences to the A3, A4, and ISO1 assemblies and retain only the hits $\geq$ 95% identity $\geq$ 95% of the query CDS length. Next, we identified genes possessing only one unique copy in ISO1 genome (i.e., every CDS in the gene has only one hit in ISO1) and two or more copies in either A3 and/or A4. If the gene

belongs to a large gene family with many similar members, they would cross-blast to each other and will return multiple hits in the ISO1-self-blast results (e.g., casein kinase 2 subunit beta family; 16 genes in that family are cross-blasted to each other). This step removes those genes with the goal of filtering out all but unique singleton genes in ISO1. At this step, we retained 62 fully duplicated genes, 12 of which are from the mitochondrial genome. We focus our attention on the 50 nuclear genes.

In addition to blasting the CDS region, we also extracted full gene in ISO1 and blasted it to A3, A4, and ISO1 itself. We retained the blast hits with $\geq 50\%$ length of the query gene length and required that the number of hits of full genes matches with the hits of CDS. For example, each CDS in the gene *Muc12Ea* has one hit while the full gene has 233 hits in ISO1-self-blast; another example is in gene *Cyp9f2*. It has one hit for the full gene blast, but two of the CDS have two hits. These genes presumably contain common motifs with other genes, which means the RNA reads from other genes may be falsely mapped to this gene by the common motifs, so we also removed these ambiguous genes. Finally, 35 genes were retained for gene expression analysis.

### 2.3.3 Measuring the gene expression in two conditions with RNA-seq

We downloaded the gene expression data from a nicotine resistance QTL mapping study carried out in the DSPR [41]. The transcriptomes were extracted from the first-instar larvae of A3 and A4 strain in two conditions (control vs. nicotine). The larvae were exposed to control media or nicotine-containing media for 4 hours before collection. The four samples (A3 control, A3 nicotine, A4 control, A4 nicotine) were then sequenced with 100 bp single-end with Illumina HiSeq 2500. We followed the paper for quality trimming with sickle (version 1.200, github.com/najoshi/sickle) and obtained 182.4 million reads for A3 control

sample, 181.4 million reads for A3 nicotine sample, 185.1 million reads for A4 control and 184.4 million reads for A4 nicotine sample.

### 2.3.4 Aggregate and independent measurement of paralog transcript abundance

We mapped A3 samples to gene sequences of the A3 assembly and A4 samples to gene sequences of the A4 assembly with hisat2 [23]. Each of the 35 genes was used as the reference individually. Take *Cyp28d1* as an example: *Cyp28d1* is a singleton in the A3 strain and has two paralogs in the A4 strain. We used the singleton sequence as reference for A3 RNA reads mapping and the two paralogs sequence as reference for A4 RNA reads mapping. We used default parameters in hisat2, except -k 20 to allow multiple mapping. We counted the reads that perfectly mapped to the reference (no SNPs or indels), and calculated the FPKM [12] for comparing expression between samples. If the paralogs of a gene have non-homologous regions, the RNA reads mapped to those regions are not considered. The paralog-specific read counts are estimated by the ratio of unique read counts mapped to each paralog. The gene that has more than two paralog copies are not considered for paralog-specific read counts estimation.

As another example, consider *Ugt86Dh*. *Ugt86Dh* is a singleton in A3 with a length of 2,384 bp. In A4, *Ugt86Dh* has two paralogs. One is of 2,005 bp, the other is of 2,382 bp including 322 bp non-homologous sequence in the end (2,060–2,382 is the non-homologous region). After the mapping, there are 1,300 reads perfectly mapped to A3 *Ugt86Dh* from the A3 control sample, whose total read counts is 182.4 million. For the A4 control sample, whose total read counts is 185.1 million, there are 1,210 reads mapped to both paralogs, 942 reads mapped to copy one uniquely and 978 reads mapped to copy two uniquely. For the second copy, 187 reads overlap with the 2,060–2,382 non-homologous region, so we don't consider

them for A4 total read counts or A4 paralog-specific read counts. The total read counts and paralog-specific read counts for A4 are calculated in the following way:

Total A4 read counts = (common read counts) + (copy1 unique) + (copy2 unique in homolog region) $= 1,210 + 942 + (978 - 187) = 2,943$;
A4 copy1 read counts = (total read counts) $\times$ (copy1 unique) / (sum of copy1 and copy2 unique in homolog region) $= 2,943 \times 942/(942 + 978 - 187) = 1,600$;
A4 copy2 read counts = (total read counts) $\times$ (copy2 unique in homolog region) / (sum of copy1 and copy2 unique in homolog region) $= 2,943 \times (978 - 187)/(942 + 978 - 187) = 1,343$.

The FPKM are then calculated in the following way:

A3 FPKM = (A3 total read counts) / (A3 length in kbp $\times$ A3 control sample total read counts in million) $= 1,300/(2.384 \text{ kbp} \times 182.4 \text{ million}) = 2.99$;
A4 FPKM = (A4 total read counts) / (max(A4 homolog region length of copy1 and copy2) $\times$ A4 control sample total read counts) $= 2,943/(\max(2.060 \text{ kbp}, 2.005 \text{ kbp}) \times 185.1 \text{ million}) = 7.72$;
A4 copy1 FPKM = (A4 copy1 read counts) / (A4 copy1 homolog region length in kbp $\times$ A4 control sample total read counts in million) $= 1,600/(2.005 \text{ kbp} \times 185.1 \text{ million}) = 4.31$;
A4 copy2 FPKM = (A4 copy2 read counts in homolog region) / (A4 copy2 homolog region length in kbp $\times$ A4 control sample total read counts in million) $= 1,343/(2.060 \text{ kbp} \times 185.1 \text{ million}) = 3.52$.

### 2.3.5 Differential expression test on total expression level and paralog-specific expression level

Since there is only one replicate of each sample, a binomial model was used for the differential expression test. One gene's raw counts for A3 and A4 were first divided by normalizing

factors to correct the difference in sequencing depth. For example, A3 nicotine sample has 181 million reads, while A4 nicotine sample has 184 million reads, then normalizing factors are 1 for A3 and 1.02 (184/181) for A4. The normalized read counts are then tested by a binomial model against a null hypothesis of equal expression.

## 2.4 Results

### 2.4.1 Distribution of the 35 paralogs selected for further study

The 35 genes we selected for additional study are scattered across all 5 major chromosome arms (Figure 2.1). As a part of our selection and filtering process, we required that the state of the ISO1 genome to be single copy. 30 of the duplication events are tandem duplicates (e.g. the three genes on chromosome arm 2L), but some tandem duplicated regions involve more than one gene (e.g. *Dlc90F* and *CG18600* on chromosome arm 3L are duplicated together). We found two genes that were duplicated to other chromosome arms (*Hapin* and *Snakeskin*). We also discovered a large duplication spanning 5 genes on chromosome arm 2L. Interestingly, the two copies are 4 million base pairs apart (Figure 2.1: chromosome arm 2R). Two of these five genes (*Ugt49B1* and *Ugt49C1*) have 5 copies in total. One possible explanation is that the two genes were duplicated tandemly before the larger duplication event affecting the whole region. These two genes duplicated again into 5 copies.

The copy number of these 35 genes ranges from 2 to 7. 15 of the 35 genes get duplicated exclusively on A3 genome, including two disperse-duplicated genes, while 19 of the 35 genes duplicated on A4 genome. The gene Ugt303B3 has 1 copy in ISO1 genome, 3 copies in the A3 genome and 2 copies in the A4 genome. Instead of being a duplication event, this phenomena could also be explained by gene deletion events, depending on the copy number of the ancestry state.

### 2.4.2 The functional distribution of the 35 genes

We performed GO analysis for the 35 genes. In terms of molecular function, the largest category is catalytic activity, which contains 15 genes (Figure 2.2a). The two largest child categories are transferase activity (6 genes: *Haspin*, *Ugt49B1*, *Ugt49C1*, *Ugt86Dh*, *Ugt303B3*, *CG1894*) and hydrolase activity (4 genes: *Pros*β*5R2*, θ*Trypsin*, *CG6472*, *Lapsyn*). The over-representation test against all genes in *Drosophila melanogaster* shows that the category "UDP-glycosyltransferase activity", a grandchild "category of transferase activity" category was over-represented with a p-value of $9e-6$. In terms of biological processes, the largest category is metabolic process, which contains 8 genes (Figure 2.2b), though no significant over-representation was shown. In terms of cellular component (Figure 2.2c), the largest category is organelle, containing 10 genes, but no significant over-representation was shown either.

### 2.4.3 Pairs of paralog genes are usually differentially expressed

16 of the 35 genes are two-copy paralogs with sufficient genetic variation between the copies to permit us to estimate their paralog-specific expression levels. We first calculated the paralog-specific FPKM and then use the raw read counts in the paralogs' homolog region to test whether they are differentially expressed (see §2.3).

The statistical significance test shows that all 16 pairs of paralogs are differentially expressed, although the magnitude varies (Figures 2.3, 2.4). The gene *spook* has no expression in one of its paralogs in either condition, suggesting that this duplication product might be a pseudogene. One copy of the gene *Lapsyn* and *CG6472* are not expressed in control conditions but are in the nicotine condition, though the read count of *Lapsyn* is low for one copy (12 for one copy, 2,930 for another copy).

Paralog pairs generally exhibit consistent expression patterns between conditions for most of the genes (the highly expressed copy in the control condition is usually still the most highly expressed in the nicotine condition). One exception is the gene *CG2233* (Figure 2.4), but the expression of the two copies do not differ much between the conditions. Three pairs of paralogs have similar expression levels (*Ugt86Dh*, *fiz*, *CG9612*) and respond similarly in control/nicotine environment, suggesting they are regulated together. There are also examples of genes where the two paralogs respond differently to environmental change. For example, for the genes *IntS3* and *mRpS5*, one copy does not show variable expression between conditions but the other shows decreased expression in the nicotine treatment (Figure 2.4).

For the two dispersed duplicated genes (*Snakeskin* and *Haspin*), the original copy is more highly expressed than the derived copy for both conditions.

### 2.4.4 Gene duplication usually leads to increased expression

When a gene is duplicated, the combined expression level of both paralogs usually increases relative to the single copy allele, but we observe some exceptions: *IntS3*, *spook*, *Dlc90F*, *Lapsyn*, *CG1894*, *Alp9*, *fiz*. For these genes, the expression levels did not change in either treatment (Figures 2.5, 2.6). From the paralog-specific expression analyses, the stable expression level for *IntS3*, *Lapsyn*, and *spook* can be explained by the observation that one copy contributed virtually nothing to the total expression level. These copies might be pseudogenes or genes with regulatory elements that prevent their expression in the treatments. Unfortunately, we cannot conclude much merely on the basis of an absence of expression evidence. For the gene *CG1894*, although the total expression level did not change in the control treatment, it did dramatically increases in the nicotine treatment (Figure 2.5: panel n=7, Figure 2.6) indicating that the six new copies in the A3 strain are regulated by elements that can be induced by nicotine. The gene *Alp9* and *fiz* are two examples where

the total expression of duplicated genes is lower the the singleton version. For *Alp9*, this is observed only in the nicotine treatment while it is observed for both treatments for *fiz*. The paralog-specific expression data shows that both copies of *fiz* in A4 strain are expressed (Figures 2.3, 2.4) suggesting that the total expression level of this gene is tightly monitored by the regulatory network in A4.

There are also genes whose expression levels are well predicted by the copy number, including: *CG31157*, *CG7966*, *Ugt49b1*, *Ugt49C1* (Figure 2.5: panel n=3, n=5, Figure 2.6, Figure 2.7). The total expression levels of the high copy-number alleles are roughly n fold higher (n=3, n=5) in both control and nicotine conditions indicating that these duplication events retained the same regulatory sequences. For the two-copy genes, the total expression levels distribute evenly around the expected 2 fold line (Figure 2.5, panel: n=2, Figure 2.7). This reflects that the total expression level is not simply proportional to its genetic content.

## 2.5 Discussion

This project is a preliminary but comprehensive survey of the effect of gene duplication on expression level. Our sampling approach required paralogs to exhibit only one unique copy in the ISO1 reference genome. For each gene in our list, the duplicate is either in A3 strain or A4 strain, the only exception is the gene *Ugt303B3*, which has 3 copies in A3 and 2 copies in A4. So the ancestral state of the gene *Ugt303B3* is unknown, but for other genes, we can infer that these are newly duplicated genes, and the duplication events happened after the A3 and A4 separation. The project does not explore the correlation of fitness and change in expression. These duplicates were observed in a small sample size and are therefore likely to be drawn disproportionately from higher frequency classes. Consequently, they might be expected to be enriched for beneficial or at least neutral variants, although we do not know their frequeny distribution among other *Drosophila melanogaster* populations.

A gene's expression level change due to duplication varies a lot in our 35 gene sample. We observed a few genes whose expression levels are roughly multiplied by the number of copies but we also observed several genes whose expression barely changes. However, we can still conclude that as long as both paralogs are expressed (i.e., they have not become pseudogenes), the expression level tends to increase, though there is subsantial variation in the realized increase in expression. In short, naive predictions based on predictions derived solely from copy number changes are frequently wrong.

The paralog-specific expression level analysis is possible only for 16 out of 35 genes due to the requirement that the paralogs must exhibit variation that uniquely identifies the members of the pairs. Frequently, the pairs are 100% identical, preventing paralog-specific expression analysis. All 16 pairs are nominally differentially expressed based on our binomial significance test. Since we only have one replicate of the expression experiment, the over-dispersion among replicate cannot be calculated. Thus the number of differentially expressed pairs we report will be an overestimate. One pattern worth mentioning is that many of the newly copied paralogs seem to have lower expression level than the original copy. When members of paralog pairs can be definitively identified as ancestral and new (e.g., dispersed-copy duplicates and three gene in the large copied region at the end of chromosome arm 2L) it is the newer copies that have the lower expression. Our catalog of incipient new genes constitutes a class of variants that merits further investigation because of their potential to advance our knowledge of regulatory architecture.

## 2.6 Acknowledgement

We thank M.Chakroborty for providing the assemblies. Yi Liao for suggestions in detecting paralogs. The work was supported by US National Institutes of Health (NIH) grant R01GM123303-1 (J.J.E.). This work was made possible, in part, through access to the High

Performance Computing Cluster of University of California, Irvine.

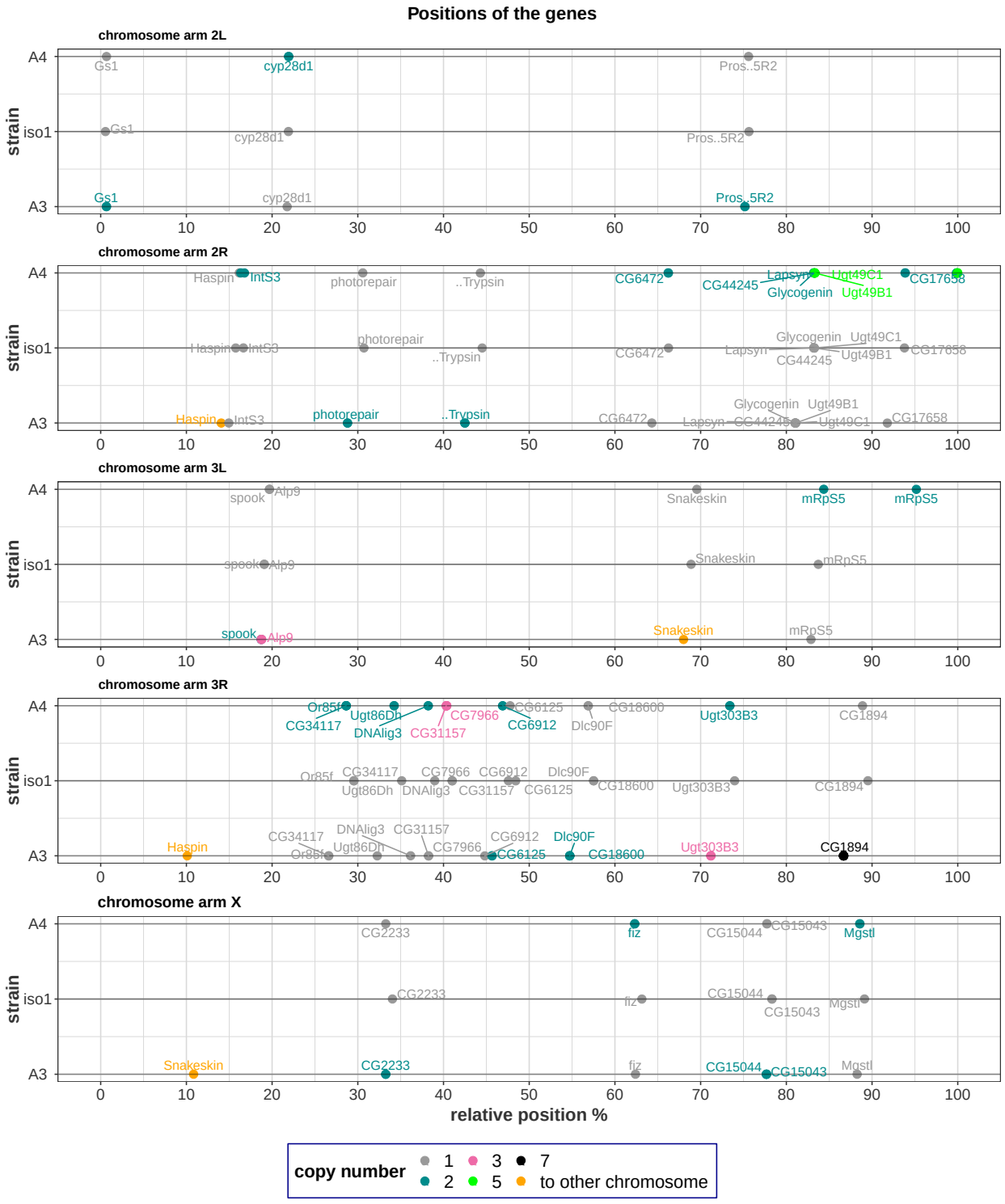



Figure 2.1: The relative position of duplicated gene in ISO1, A3, A4 strains are plotted. The color of dots and gene names shows the gene copy number in each strain. Some genes are located together and also duplicated together; Each tandem are shown by only one dot but the gene names are clearly labeled.

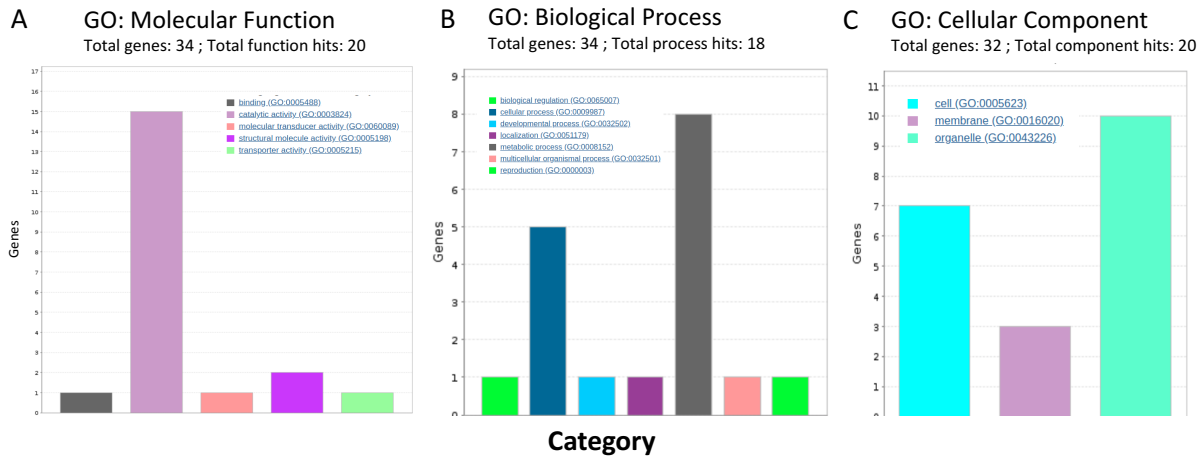



Figure 2.2: The 35 genes are classified by Molecular Function, Biological Process and Cellular Component. The fisher exact test for over-representing shows the "UDP-glycosyltransferase activity" (a category belongs to "catalytic activity") is significantly over-represented among all genes in *Drosophila*, but no significant over-representing was shown for Biological Process or Cellular Component.

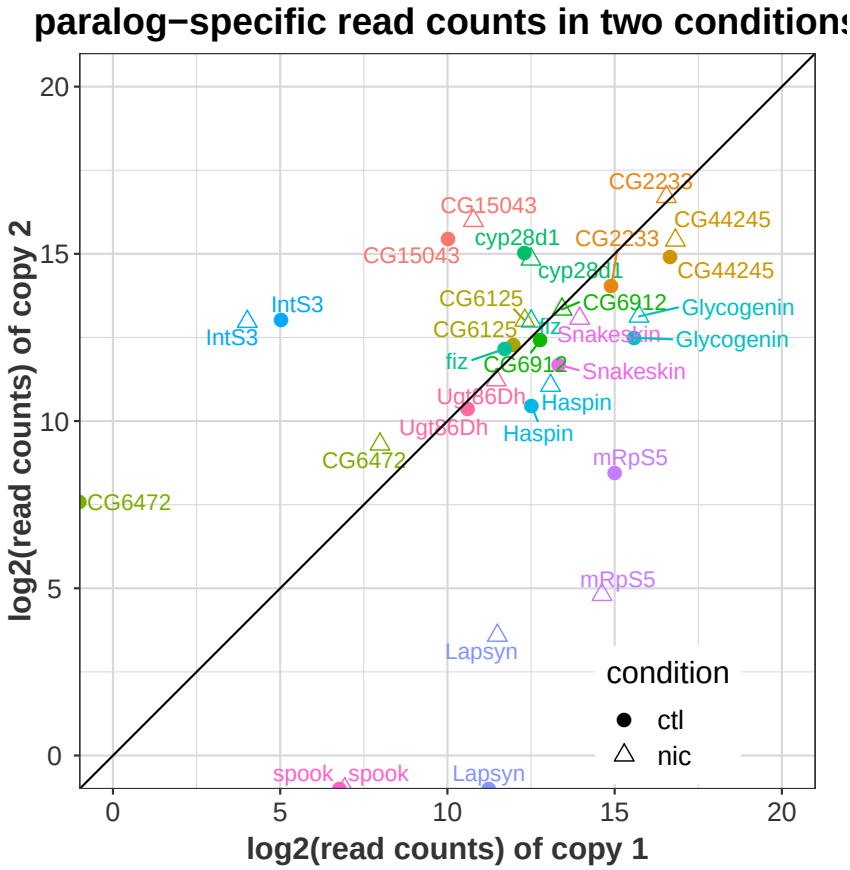



Figure 2.3: 16 genes' paralog-specific read counts in control and nicotine conditions are plotted. The read counts are plotted on log space. The color of the gene names made it easier to identify the same gene of two conditions. The diagonal line indicates equal expression of two paralogs. The statistical significance test shows that all 16 pairs of paralogs express differently in both conditions, although the magnitude varies (the distance to the diagonal line).

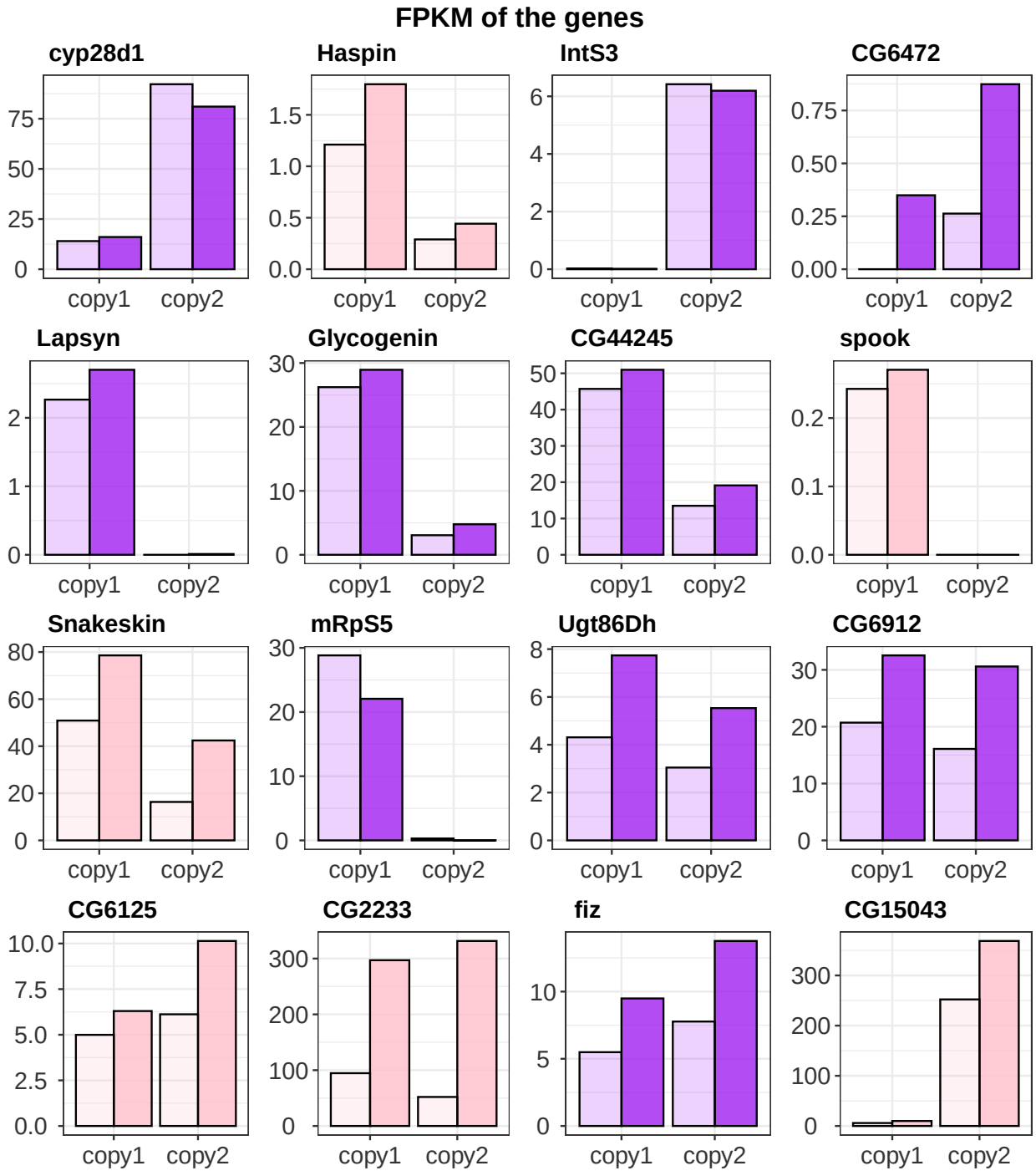



Figure 2.4: The color indicates the strain in which duplication occurs. Purple means duplication occurs in A4 strain, pink means duplication occurs in A3 strain. Few genes has small FPKM, which are less convincing for expression level. However, their raw read counts are sufficient to made qualitative conclusions. The highest bin of Haspin, CG6472, spook has raw read counts of 8672, 642, 122 respectively.

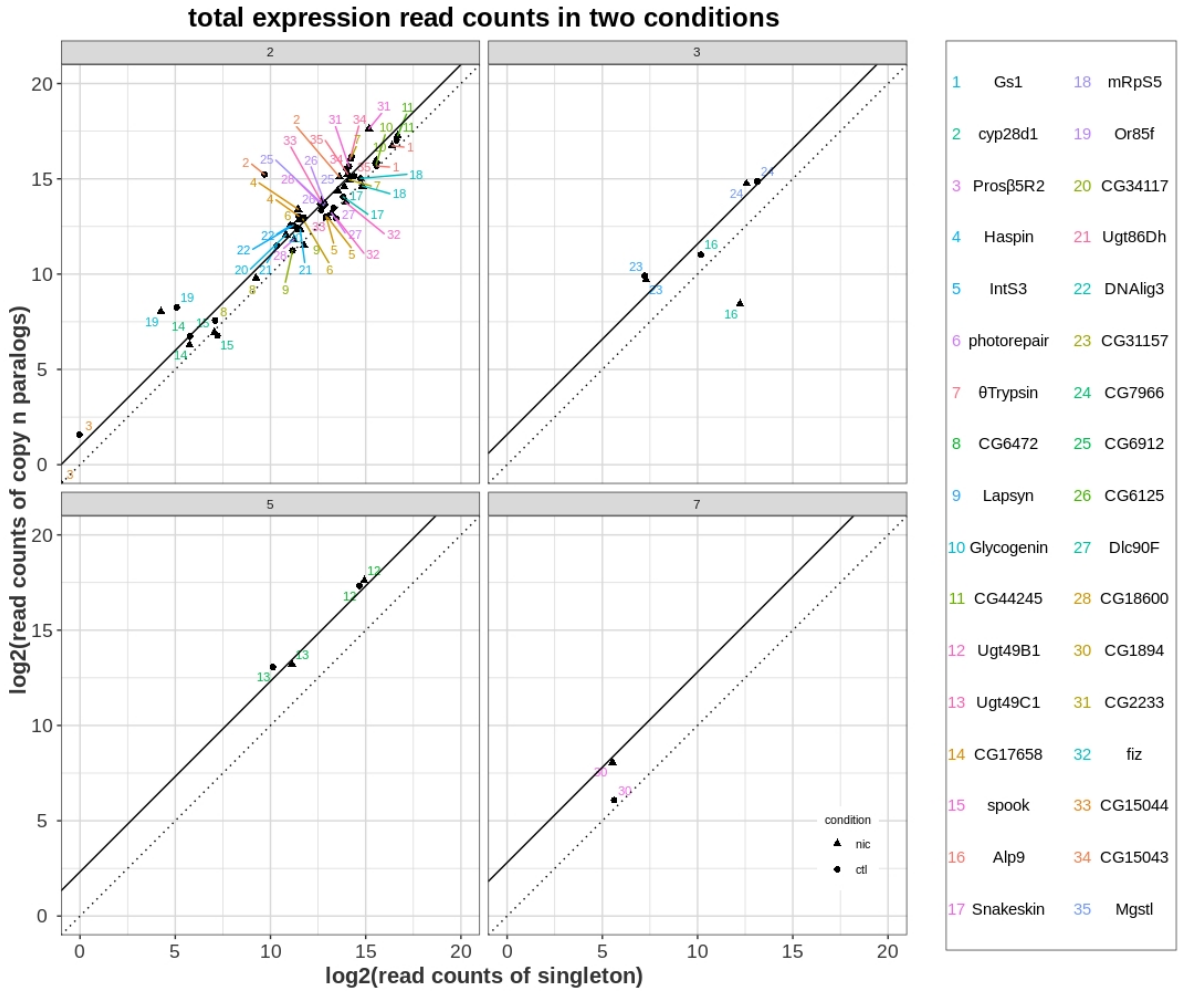



Figure 2.5: The raw read counts in A3 and A4 strain are plotted. The singletons' expression is on x axis(no matter in A3 or A4), and the total expression of the multi-copy version is on y axis. The number on top of each panel is the copy number. The dotted line in each panel is the equal express line. The solid line indicates the expected relationship of expression level if completely conform to the copy number.

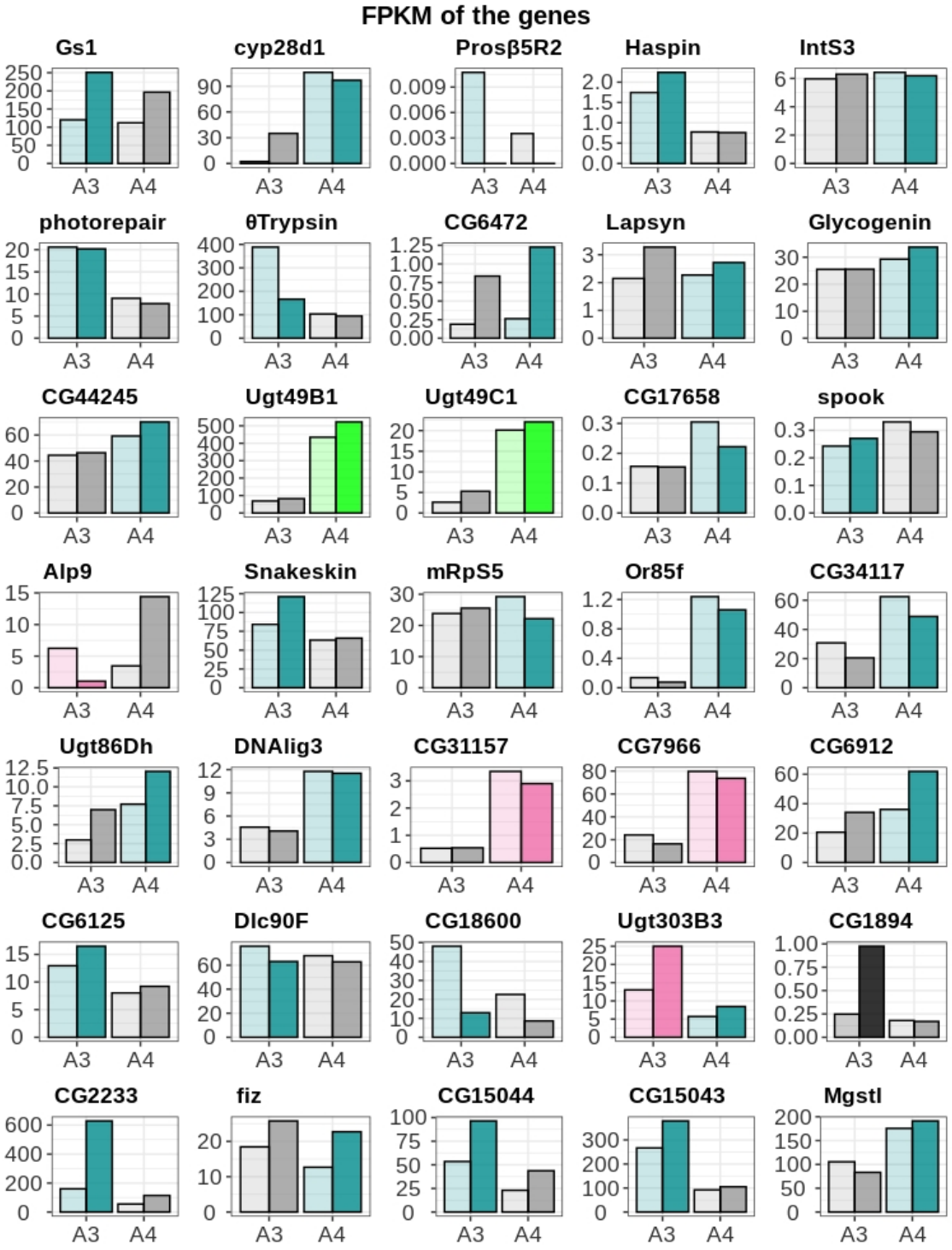



Figure 2.6: The figure shows the FPKM of the total expression level in 35 genes. The color scheme is the same as in figure 1. Some genes have small FPKM (less than 1), but their raw reads may be sufficient to make qualitative conclusions. The raw read counts of the highest bin of these genes are listed here: Pros$\beta$5R2:3; CG6472:900; CG17658: 109, spook: 152; Or85f: 312; CG1894: 266;. The raw read counts shows that Pros$\beta$5R2 is barely expressed, the expression level of it is not reliable.

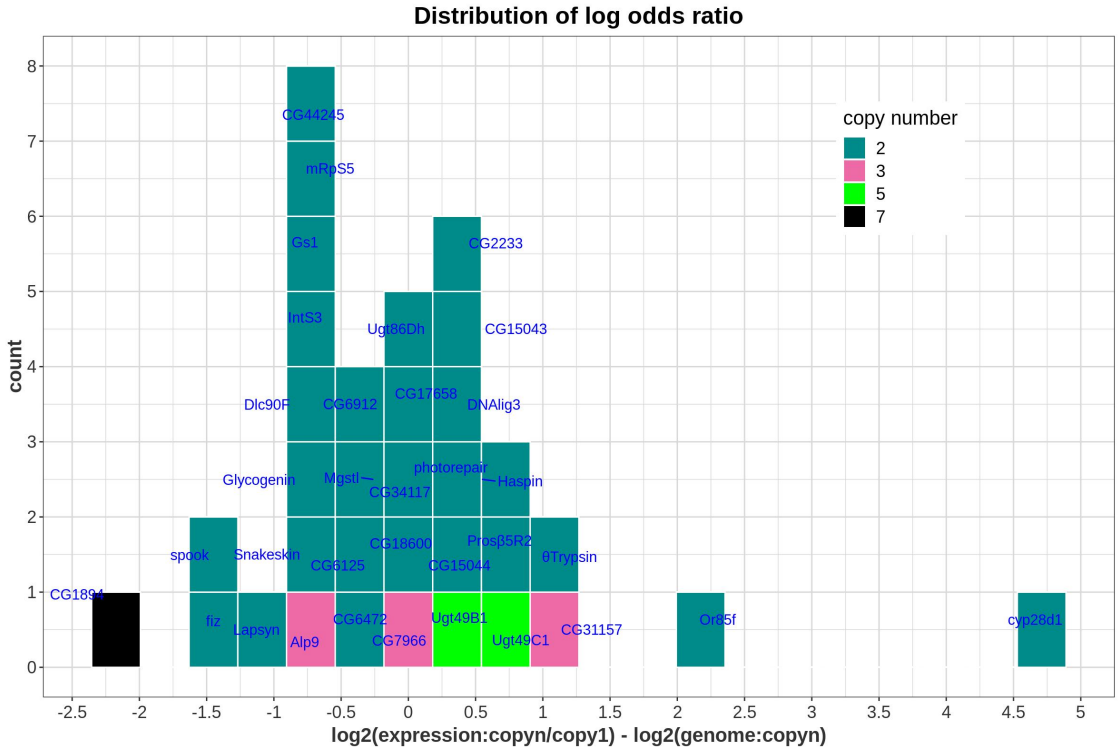



Figure 2.7: The figure shows the distribution of log odds ratio of 35 genes in control condition. The log odds ratio is calculated as the log2 fold change in expression minus the log2 fold change in gene copy. The color of the block uses the same color pattern in the FPKM graphs. The gene names is also labeled in each block.

# Chapter 3

# Inferring Genetic Architecture of Expression Variation

In the project, I described a new beta-binomial model for general *cis* variation measurement and applied to 20 replicated yeast hybrid samples. Along the process, the common mapping bias problem was also solved by a carefully-designed pipeline. This work is preprinted in bioRxiv under the title "Inferring the genetic architecture of expression variation from replicated high throughput allele-specific expression experiments". It is reprinted here with a few modifications to conform the dissertation format.

## 3.1 Abstract

Gene expression variation between alleles in a diploid cell is mediated by variation in *cis* regulatory sequences, which usually refers to the differences in DNA sequence between two alleles near the gene of interest. Expression differences caused by *cis* variation has been estimated by the ratio of the expression level of the two alleles under a binomial model.

However, the binomial model underestimates the variance among replicated experiments resulting in the exaggerated statistical significance of estimated *cis* effects and thus many false discoveries of *cis*-affected genes. Here we describe a beta-binomial model that estimates the *cis*-effect for each gene while permitting overdispersion of variance among replicates. We demonstrated with simulated null data (data without true *cis*-effect) that the new model fits the true distribution better, resulting in approximately 5% false positive rate under 5% significance level in all null datasets, considerably better than the 6%–40% false positive rate of the binomial model. Additional replicates increase the performance of the beta-binomial model but not of the binomial model. We also collected new allele-specific expression data from an experiment comprised of 20 replicates of a yeast hybrid (YPS128/RM11-1a). We eliminated the mapping bias problem with de novo assemblies of the two parental genomes. By applying the beta-binomial model to this dataset, we found that *cis* effects are ubiquitous, affecting around 70% of genes. However, most of these changes are small in magnitude. The high number of replicates enabled us a better approximation of *cis* landscape within species and also provides a resource for future exploration for better models.

## 3.2 Introduction

Variation in gene expression contributes significantly to phenotypic variation [22, 42]. Consequently, gene regulatory elements have long been thought to be an important target of natural selection comparable in significance to variation in the proteome [47, 26, 76]. The genetic architecture of variation in gene regulation can be decomposed into *cis* variation and *trans* variation. The *cis* variation affects expression differences between two individuals in a non-diffusible manner (e.g., a mutation on a promoter region), while *trans* variation affects the expression difference in a diffusible manner (e.g., a coding region mutation on a transcription factor) [15, 74].

In an F1 hybrid cell, two alleles of the same gene are exposed to the same diffusible elements, so any difference between the alleles' expression must be encoded by features linked to the gene itself (i.e., *cis* variation). By measuring the allele-specific expression of all genes in hybrid cells, we can measure the magnitude of *cis* variation (*cis*-effect) and detect *cis*-affected genes [62, 15]. The *cis* effect parameter ($e_{cis}$) for a gene is defined as the ratio of the expression from allele 1 and allele 2 [14, 58]. However, previous allele-specific expression studies using RNA-seq for *cis*-effect typically employed 1-3 hybrid replicates in binomial framework [14, 58, 45, 52, 39, 44, 3], which assumes that the read counts for each allele among replicates can be modeled as a Poisson random variable.

The actual variance among RNA-seq experiments is known to be overdispersed, and consequently, the single Poisson parameter is inadequate to model both the mean and variance. The negative binomial distribution instead has been shown to fit better than Poisson in many differential expression studies [53, 59, 18]. The negative binomial distribution is equivalent to the compound gamma-Poisson distribution, where the lambda parameter of Poisson is a gamma-distributed random variable. The two parameters of the negative binomial permit the mean and variance to vary independently. Therefore, we modeled allelic expression for each gene with a negative binomial distribution instead of a Poisson distribution. Under this assumption, the *cis*-effect $e_{cis}$ is beta-binomially distributed with an overdispersion parameter compared with the binomial distribution (see §3.3.6).

We compared the false positive rates of the two models with simulated null datasets where no true *cis* effects exist. We found that the binomial model has high false positive rate even with a large number of replicates, but the beta-binomial model improves with increased replication, attaining a 5% false positive rate as expected.

We also grew 20 replicates of hybrid from the cross of yeast Saccharomyces cerevisiae strains YPS128 and RM11-1a to estimate *cis* variation. YPS128 is a woodland stain [65] and RM11-1a is a derivative of a vineyard strain [6]. We used RNA-seq for allele-specific counts and

estimated the gene-wise $e_{cis}$ with both models. In terms of power, both models improve as replication increases. We found from this experimental data that ∼70% of the total 4,710 informative genes have a significant *cis* difference. Around 2% of the total genes have a greater than 2-fold difference significantly.

Estimated from the simulated null data, 20%–30% genes lacking a true *cis* effect would be falsely classified as significant by the binomial model. In our experimental data, the beta-binomial model and binomial model differ by ∼5% in the number of significant *cis* affected genes (Figure 3.4), which is less than the 15%–25% difference in false-positive rate estimated from the null data. This could perhaps be explained by the possibility that the two strains are sufficiently diverse that most of the genes are true positives. However, for closely related species (or strains) with less differential gene expression, a 5% false positive rate would contribute a much higher proportion to the total number of differentially expressed genes.

This allele-specific study demonstrated the advantage of the beta-binomial model over the binomial model and the salutary effect of using high replication. The high number of replicates of hybrid samples between the two yeast strains enabled us a better approximation of *cis* landscape within species. It also provides a resource for future exploration of better models.

## 3.3 Materials and Methods

### 3.3.1 Yeast strains and preparation of hybrid samples

We mixed a single colony of YPS128 strain (MAT$\alpha$; ura3::kanMX; HO::HygMX; lys2::ura3) and a single colony of RM11-1a strain (MATa; leu2$\Delta$0; ura3$\Delta$0; HO::kanMX) together in 100 ul YPAD, put the mix in 30 °C for 4 hours, then we poured 50ul of mixed cells into a

dropout plate (-leu, -lys), and struck to get one single diploid colony.

We picked one single diploid colony and struck it on the standard YPAD plate for hybrid sample collection. We then collected 20 independent hybrid samples started from this YPAD plate. Each sample was generated by the following procedure:

One single colony was taken from the YPAD plate and was cultured overnight. It was then diluted to OD 0.05 in 5 ml YPAD and grow until OD 0.7–0.8 in 30 °C with 220 rpm shaking. The yeast culture was then distributed in Eppendorf tubes by 1 ml per tube, centrifuged with 9,000 rpm to remove the supernatant, snap-frozen in liquid Nitrogen and finally stored at −80 °C for DNA and RNA extraction.

### 3.3.2 DNA extraction and sequencing

We extracted the DNA of these 20 hybrid samples using the Yeast DNA Extraction Kit (Thermo Scientific 78870). After extraction, we used the Nextera DNA Library Preparation Kit (Illumina) to make 20 libraries with unique barcode combination (Nextera Index Kit) and pooled them together before sequencing. We sequenced the pooled library in UC Davis Genome Center (`http://dnatech.genomecenter.ucdavis.edu/`) with 1 Lane of mid-output Nextseq PE75. We then demultiplexed [51] the pooled reads and got a total of 95.3 million reads for the 20 replicates.

### 3.3.3 RNA extraction and sequencing

We extracted the RNA of these 20 hybrid samples using the TRIzol Plus RNA Purification Kit (Invitrogen). Transcriptome libraries were made by the Smart-seq2 protocol [49]. The 20 Libraries were pooled together and sequenced in UC Davis Genome Center with 4 Lanes of high-output Nextseq PE75. After demultiplexing [51], we got a total of 1,530.8 million

reads for the 20 replicates.

### 3.3.4 Sequencing and assembly of YPS128 and RM11-1a genome

We assembled our YPS128 and RM11-1a genome and used them as the reference genomes in mapping DNA/RNA reads. We extracted whole genome DNA of YPS128 strain and RM11-1a strain using the QIAamp DNA Mini Preparation Kit (Qiagen), prepared Nextera DNA library (Illumina) and sequenced the pooled library with 1 Lane of Miseq PE75, which generated an 88X coverage for the YPS128 strain and a 102X coverage for the RM11-1a strain. We also generated long DNA reads with Oxford Nanopore (Rapid sequencing) for RM11-1a strain and got a 59X coverage. Since our Nanopore experiment failed for the YPS128 strain, we downloaded its Pacbio long reads from this project [77] which gives a 230X coverage.

For YPS128 strain, we used Dextractor (`https://github.com/thegenemyers/DEXTRACTOR`) to extract fastq sequences from the original h5 files. Then we used Canu [29] for raw assembly and finisherSC [31] for gap fixing, followed by two rounds of quiver (`https://github.com/PacificBiosciences/GenomicConsensus`) correction. We further polished the assembly with Illumina short reads using pilon [70] and pacbio long reads again using quiver followed by one final round of pilon. We ended up with an assembly with NG50=808.6K and Busco score [63] of 94.4% (fungi).

For RM11-1a strain, we used Albacore (ONT software version 2.2.7) for nanopore long reads base-calling. Then we used Canu [29] for raw assembly followed by finisherSC [31] for gap fixing, then corrected the raw assembly by three rounds of Racon (`https://github.com/isovic/racon`). We further polished the assembly with the Illumina short reads using Pilon [70] and nanopore long reads again using Racon. We did the pilon-racon for two rounds and wrapped up with four rounds of pilon. Finally, we obtained an assembly with NG50=919.8K

and Busco score [63] of 93.7% (fungi).

The qualities of the assemblies are further evaluated with QV estimation. We aligned the Illumina reads used for polishing to the final assembly using bwa mem [35] with default parameters. Following [28], we used freebayes (V.1.2.0-4) [17] to estimate the number of SNPs and indel variants with the command `freebayes -C 2 -0 -O -q 20 -z 0.10 -E 0 -X -u --ploidy 1 -F 0.75 -f asm.fasta asm_nodup.bam > asm.vcf`. Total bases changed E (inserted, deleted, substituted) was summed and divided by the total number of bases (T) with minimum coverage 3. QV was calculated as $-10\log_{10}(E/T)$.

### 3.3.5 Collect DNA/RNA read counts

#### Identify variants between YPS128 and RM11-1a

The reads from hybrid samples are unidentifiable of which parental genotype they belong to if they do not overlap with any variant (SNPs or Indels) between the two parental strains. So we first extracted a list of SNPs and Indels by comparing the YPS128 assembly and RM11-1a assembly using MUMmer [30] (MUMmer/3.23: nucmer; show-snps). For conservativeness, we did it in both directions (using YPS128 as query, RM11-1a as subject and then exchange) and only retained the SNPs and Indels that appear in both comparisons.

#### Mapping DNA reads with two references

We next mapped the DNA reads of the 20 hybrid samples to both assemblies using bowtie2 [32] (bowtie2.2.7) and got 40 mapping files. We then counted the allele-specific number of reads hitting each variant position with Samtools [36] (Samtools 1.9: mpileup setting -q to 5 to ignore multi-hits reads) and customized scripts (`count_pileup.py`: count the number of reads mapping to the reference allele and alternative allele respectively using mpileup output

file as input).

We found that the mapping always biases towards the reference genome. In hybrid DNA samples, the reads from YPS128 genome is expected to be of the same amount as from RM11-1a genome. However, when YPS128 assembly was used as the reference genome, the sum of reads assigned to YPS128 allele across all variants is around 1.4 fold more than the sum of reads assigned to RM11-1a allele in all of the 20 hybrids. This also happened when RM11-1a was used as the reference genome. The sum of reads assigned to RM11-1a allele across all variants is around 1.4 fold more than the number of reads assigned to YPS128 allele (Figure 3.7).

One main reason for this mapping bias is that when one assembly was chosen as the reference genome, the reads from the alternative genome in the hybrid sample are not as likely to map to the correct genomic position because of the variant. So we conceived that the alternative counts for each SNP/indel in the mapping results are underestimated while the reference counts are more reliable. Thus, we only kept the YPS128 allelic reads from mapping results using YPS128 as the reference genome and RM11-1a allelic reads from mapping results using RM11-1a as the reference genome. Some variant positions are close to each other and the reads that cover both of them would be counted repeatedly when summing up the counts, so we also unioned the reads from each allele using the reads' names as identifiers. After this operation, we reduced most of the mapping bias, but the total read counts still biased towards RM11-1a genome by around 1% (Figure 3.8).

**Identifying suspected loci causing mapping bias**

Another possible source for mapping bias are the errors in genome assembly and the in-coordination between the assembly and the real genotype in hybrid (the YPS128 strain's Pacbio long-reads used in assembly is not from this project), or regions that the sequencing

probability for two alleles is extremely different. In these loci, nearly all the allelic read counts would be assigned to one of the genomes. As these kinds of loci accumulate, the bias would be reflected in the total read counts. Thus, we check the reads that cover each variant position to see whether the nucleotide information provided by the short reads in hybrid samples match with the variant we got from genome comparison. For example, If the SNP pair is A on YPS128 and C on RM11-1a from the comparison of assemblies, short reads with A and short reads with C on the corresponding positions are both required to exist in all mapping results. Variants without sufficient short reads support were removed for downstream analysis (12,793 positions are removed from total 82,029 positions in YPS128; 12,326 positions are removed from total 81,574 positions in RM11-1a). After the removal of those positions, we recounted the read counts overlapping with the remaining positions and also unioned the reads covering consecutive positions (`group_reads.py`, `yps5rmB_gc.py`), the mapping bias was then sufficient small to be ignored (Figure 3.9).

**Mapping expression reads and collecting allele-specific read counts**

We first annotated the two assemblies with CrossMap [80] v0.2.8 (using S.cer reference annotation), and label each variant position with gene name. The variant positions that are not in any gene regions or overlap with two gene range (Some gene overlaps in yeast) are further removed. There are 37,487 variant positions retained, which cover 4,710 genes.

We then mapped the Expression reads using bowtie2 (bowtie 2.2.7: there are very limited intron regions in the yeast genome, so we did not choose an RNA splice-sites aware mapping tool) to both two assemblies. Same as the procedure for DNA reads counting, we collected counts from only reference allele for each retained position.

Finally, we aggregated the read counts of variant positions under the same gene name and counted allelic reads with samtools and customized scripts as we did for DNA counts

(`group_reads.py`, `yps5rmB_gc.py`). The total read counts of YPS128 allele and RM11-1a allele are almost the same in the 20 hybrid samples (Figure 3.10).

#### Remove bad replicates

We checked the correlation of the read counts between each of the 40 allele-specific expression profiles (function `cor()` in R (R Foundation for Statistical Computing, Vienna, Austria., n.d.), Figure 3.11), and found that the expression profiles from two replicates 14A and 9A are apparently different from other replicates. These two replicates are happened to be the two outliers in Figure 3.10 (the leftmost and rightmost point). We decided to remove them for downstream analysis.

### 3.3.6 *Cis* variation estimation

We use the ratio of two alleles' expression in the hybrid to measure *cis* variation $e_{cis}$ between the two alleles in one gene.

#### Binomial model

If one assumes that the read counts in a gene for two alleles X and Y in one sample can be modeled by independent Poisson Variables, X and Y can be expressed as:

$$\begin{aligned} X &\sim Pois(\mu_1 = C_1 \cdot \lambda_1) \\ Y &\sim Pois(\mu_2 = C_2 \cdot \lambda_2) \end{aligned} \tag{3.1}$$

$C_1$ represents the total read counts from one genotype which X allele rested on. $C_2$ represents the total read counts from the other genotype which Y allele rested on. In true hybrid

samples, $C_1$ and $C_2$ are almost the same, but in simulations or parental samples they are not necessarily the same; $\lambda_1$ and $\lambda_2$ represent the proportion of reads mapping to the corresponding alleles. The total read counts $C_1$ and $C_2$ are variable across biological replicates, while $\lambda_1$, $\lambda_2$ are assumed to be biological properties of a gene (expression level) that keep constant across biological replicates. The *cis* effect ($e_{cis}$) is related to the mapping rate parameter $\lambda$ as follows:

$$e_{cis} = \frac{\lambda_1}{\lambda_2} \tag{3.2}$$

Conditionally on $X{+}Y{=}n$, the probability of $k$ reads mapped to X allele ($X{=}k$) is:

$$\begin{aligned}
&P(X{=}k|X{+}Y{=}n) \\
&= \frac{P(X{=}k \wedge X{+}Y{=}n)}{P(X{+}Y{=}n)} = \frac{P(X{=}k) \cdot P(Y{=}n{-}k)}{P(X{+}Y{=}n)} \\
&= \frac{e^{-\mu_1} \cdot \frac{\mu_1^k}{k!} \cdot e^{-\mu_2} \cdot \frac{\mu_2^{n-k}}{(n-k)!}}{e^{-(\mu_1+\mu_2)} \cdot \frac{(\mu_1+\mu_2)^n}{n!}} \\
&= \frac{n!}{k! \cdot (n-k)!} \cdot \left(\frac{\mu_1}{\mu_1+\mu_2}\right)^k \cdot \left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{n-k} \\
&= \binom{n}{k} \cdot \left(\frac{C_1{\cdot}\lambda_1}{C_1{\cdot}\lambda_1 + C_2{\cdot}\lambda_2}\right)^k \cdot \left(\frac{C_2{\cdot}\lambda_2}{C_1{\cdot}\lambda_1 + C_2{\cdot}\lambda_2}\right)^{n-k}
\end{aligned} \tag{3.3}$$

So, the read counts of X allele can be modeled by a binomial distribution conditioned on the sum of the two alleles:

$$\begin{aligned}
&X|X{+}Y{=}n \sim B(n,p) \\
&p = \frac{C_1{\cdot}\lambda_1}{C_1{\cdot}\lambda_1 + C_2{\cdot}\lambda_2} = \frac{C_1{\cdot}e_{cis}}{C_1{\cdot}e_{cis} + C_2}
\end{aligned} \tag{3.4}$$

The pdf (probability density function) for X allele's count in one sample is

$$f(X{=}k, X{+}Y{=}n, e_{cis}) = \binom{n}{k} \cdot \left(\frac{e_{cis}}{e_{cis}+1}\right)^k \cdot \left(\frac{1}{1+e_{cis}}\right)^{n-k} \tag{3.5}$$

Since the reads count variable X is independent across t biological replicates, the joint pdf is the product of the above pdf. Thus, we can use the Maximum likelihood method to estimate ecis. The log-likelihood function to maximize is

$$l(e_{cis}\ ;\ k_i, n_i, C_{1i}, C_{2i}) = \sum_{i=1}^{t} \ln f(k_i, n_i, C_{1i}, C_{2i}|e_{cis}) \tag{3.6}$$

For accommodating the `mel2()` function in R, in which we applied the log-likelihood function, the optimization for $e_{cis}$ is done on log space. The output is $\log_2(e_{cis})$ and its confidence interval.

**Beta-binomial model**

The assumption that the read counts for alleles can be modeled by independent Poisson Variables may not be appropriate since there is usually more variability than the Poisson Model.

The negative-binomial model provides a good fit to the gene-level read counts distribution [53]. It is equivalent to the gamma-Poisson model where the Poisson rate is gamma distributed, adding one degree of freedom to adjust the variance independently of the mean. We now use negative-binomial variables to model the read counts mapped to allele X and Y in the hybrid sample.

$$\begin{aligned} X &\sim NB(C{\cdot}r_1, p) \\ Y &\sim NB(C{\cdot}r_2, p) \end{aligned} \tag{3.7}$$

The mean, variance and variance-to-mean ratio for X and Y are shown below:

$$\begin{aligned}
\mathbb{E}(X) &= C_1 \cdot \lambda_1 = \frac{C_1 \cdot r_1 \cdot p}{1-p} \\
\mathrm{Var}(X) &= \frac{C_1 \cdot r_1 \cdot p}{(1-p)^2} \\
\mathbb{E}(Y) &= C_2 \cdot \lambda_2 = \frac{C_2 \cdot r_2 \cdot p}{1-p} \\
\mathrm{Var}(Y) &= \frac{C_2 \cdot r_2 \cdot p}{(1-p)^2} \\
\text{Variance to Mean Ratio: } D &= \frac{1}{1-p}
\end{aligned} \tag{3.8}$$

$C_1$ and $C_2$ represent the total read counts for each genotype in the sample as in binomial model; $\lambda_1$ and $\lambda_2$ represent the proportion of reads mapping to the corresponding alleles; $\lambda_1 = r_1 \cdot p/(1-p)$ and $\lambda_2 = r_2 \cdot p/(1-p)$. The assumption for the above modeling is that the two alleles of the same gene have the same variance-to-mean ratio $D$ ($p$ is a constant for X and Y). It is necessary for deriving the beta-binomial distribution below. Although this assumption can not reflect reality completely, it is still more relaxed than the previously used Poisson model in which the variance equals the mean. When $p$ approaches 0, the negative-binomial model approaches the Poisson Model.

The *cis* variation $e_{cis}$ is related to the parameter $r$ in the above model:

$$e_{cis} = \frac{r_1 \cdot p/(1-p)}{r_2 \cdot p/(1-p)} = \frac{r_1}{r_2} \tag{3.9}$$

Conditionally on $X{+}Y{=}n$, the probability of $k$ reads mapped to X allele ($X{=}k$) is:

$$
\begin{aligned}
& P(X{=}k|X{+}Y{=}n) \\
&= \frac{P(X{=}k \wedge X{+}Y{=}n)}{P(X{+}Y{=}n)} = \frac{P(X{=}k) \cdot P(Y{=}n{-}k)}{P(X{+}Y{=}n)} \\
&= \frac{\frac{\Gamma(k+C_1 \cdot r_1)}{k! \cdot \Gamma(C_1 \cdot r_1)} \cdot p^k \cdot (1-p)^{C_1 \cdot r_1} \cdot \frac{\Gamma(n-k+C_2 \cdot r_2)}{(n-k)! \cdot \Gamma(C_2 \cdot r_2)} \cdot p^{n-k} \cdot (1-p)^{C_2 \cdot r_2}}{\frac{n+C_1 \cdot r_1+C_2 \cdot r_2}{n! \cdot \Gamma(C_1 \cdot r_1)+C_2 \cdot r_2} \cdot p^n \cdot (1-p)^{C_1 \cdot r_1+C_2 \cdot r_2}} \\
&= \frac{\Gamma(k+C_1 \cdot r_1) \cdot \Gamma(n-k+C_2 \cdot r_2) \cdot \Gamma(C_1 \cdot r_1+C_2 \cdot r_2)}{\Gamma(C_1 \cdot r_1) \cdot \Gamma(C_2 \cdot r_2) \cdot \Gamma(n+C_1 \cdot r_1+C_2 \cdot r_2)} \cdot \frac{n!}{k! \cdot (n-k)!} \\
&= \binom{n}{k} \cdot \frac{B(k+C_1 \cdot r_1, n-k+C_2 \cdot r_2)}{B(C_1 \cdot r_1, C_2 \cdot r_2)}
\end{aligned} \tag{3.10}
$$

So, the read counts of X allele can be modeled by a beta-binomial distribution conditioned on the sum of the two alleles:

$$X|X{+}Y{=}n \sim BetaBinomial(k, n, C_1 \cdot r_1, C_2 \cdot r_2) \tag{3.11}$$

In order to incorporate $e_{cis}$ into the distribution, we reparametrize the beta-binomial distribution with $e_{cis}$ and $\theta$ which describes the over-dispersion of the beta-binomial distribution from the corresponding binomial distribution. Let:

$$\theta = \frac{1}{r_1 + r_2} \tag{3.12}$$

Then from Equation (3.8):

$$\theta = \frac{1}{r_1 + r_2} = \frac{p}{(1-p) \cdot (\lambda_1 + \lambda_2)} \tag{3.13}$$

It shows that $\theta$ is positively correlated with $p$. Then, together with Equation (3.9), we got:

$$\begin{aligned} r_1 &= \frac{e_{cis}}{\theta{\cdot}(e_{cis}+1)} \\ r_2 &= \frac{1}{\theta{\cdot}(e_{cis}+1)} \end{aligned} \tag{3.14}$$

The beta-binomial model approaches the binomial model when $\theta$ approaches zero. With the new parameterization, the pdf (probability density function) for X allele's count in one sample is

$$\begin{aligned} & f(X{=}k, X{+}Y{=}n, e_{cis}, \theta) \\ &= \binom{n}{k} \cdot \frac{B(k + C_1{\cdot}r_1, n-k+C_2{\cdot}r_2)}{B(C_1{\cdot}r_1, C_2{\cdot}r_2)} \\ &= \binom{n}{k} \cdot \frac{B(k + C_1{\cdot}\frac{e_{cis}}{\theta{\cdot}(e_{cis}+1)}, n-k+C_2{\cdot}\frac{1}{\theta{\cdot}(e_{cis}+1)}}{B(C_1{\cdot}\frac{e_{cis}}{\theta{\cdot}(e_{cis}+1)}, C_2\frac{1}{\theta{\cdot}(e_{cis}+1)})} \end{aligned} \tag{3.15}$$

Since the reads count variable X is independent across t biological replicates, the joint pdf is the product of the above pdf. Thus, we can use the Maximum likelihood method to estimate the *cis* variation $e_{cis}$ along with the over-dispersion parameter $\theta$. The final log-likelihood function to maximize is

$$l(e_{cis}, \theta \ ; \ k_i, n_i, C_{1i}, C_{2i}) = \sum_{i=1}^{t} \ln f(k_i, n_i, C_{1i}, C_{2i} | e_{cis}, \theta) \tag{3.16}$$

As in the binomial model, the final estimation of ecis, $\theta$ and their confidence intervals are on log space. The outputs are $\log_2(e_{cis})$ and $\log_2(\theta)$ and their confidence intervals.

### $C_1$ and $C_2$ parameter estimation

For calculating the $e_{cis}$ for a gene, the maximum likelihood method for both models need 4 input from each replicates: $k_i$, $n_i$, $C_{1i}$, $C_{2i}$.

$C_{1i}$ and $C_{2i}$ are the total expression read counts of the two genotypes. Since there are around 80% of reads in hybrid samples cannot be identified of which genome they belong to, the total allelic reads number cannot be known accurately.

Here we just used the total identifiable read counts from YPS128 allele as $C_1$ and those from RM11-1a allele as $C_2$ for each sample. That is to say that the aforesaid $\lambda_1$ and $\lambda_2$ are no longer the mapping rate relative to total allelic read counts but to total identifiable allelic read counts. This does not affect the estimation of $e_{cis}$ and its confidence interval If we assume that the identifiable read counts are proportional to true read counts of each allele in the hybrid samples.

### 3.3.7 Generate null datasets lacking *cis*-variation

In order to compare the binomial model and the beta-binomial model. We generated two datasets from experimental data which in principle should have no *cis*-variation and four datasets from negative-binomial (gamma-Poisson) distributed random number.

#### Null datasets from experiments

The first dataset "`Gier2015`" was generated from a haploid yeast gene expression study [18] which has 48 biological replicates under the same condition: Δsnf2. We downloaded the short reads data from ENA (ENA archive, Project ID: PRJEB5348), then, as described in the paper, got rid of four bad replicates (rep6, rep13, rep25, rep35) and obtained gene read counts with TopHAT2 [24] and HTseq [1]. We then combined every two haploid expression profiles into 1,892 ($P(44, 2) = 44\times43$) hybrid samples (Table 3.1).

The second dataset "`Xinw2018_yps`" was generated in a similar way but from our hybrid samples. We combine every two gene expression profiles from 18 qualified replicates of

YPS128 allele, which generated 306 ($P(18,2) = 18\times17$) no-*cis* hybrid samples (Table 3.1).

Some simulated hybrid samples have less variation between two alleles, some have more, but by doing this permutation and the bootstrap (see below), the structural bias from choosing extreme hybrids by chance can be attenuated and the average effect of models can be obtained.

**Null datasets from random number**

Although the null dataset generated from experimental data should in principle have no *cis*-variation, the variation between alleles is not controlled and the true underlying distribution is unknown. So to test both binomial and beta-binomial model with fully-defined hybrid samples, we generated four datasets "`simu_null:1-4`" for 5,000 genes from the negative-binomial (gamma-Poisson) distribution (Table 3.1).

The expression counts of each allele for gene $i$ was generated from the negative-binomial distribution using R:

$$X_i \sim NB(C{\cdot}r_i, p_i)$$

We set $C = 1e6$; $p_i$ was set to 0.1, 0.4, 0.8 respectively for "`simu_null:1-3`". For "`simu_null` ↪ `:4`", we used a variable $p$ for each gene, which was chosen randomly from a uniform distribution of $(0, 0.8)$. We made the gene $i$ have the same expected mapping probability $\lambda$ across the four datasets, which was chosen by randomly picking a gene from our experimental expression data and use its averaged mapping probability. Since the mapping probability for each gene is set, $r_i$ for each gene was then calculated ($r_i = \lambda{\cdot}(1-p)/p$) and used as a parameter to generate $X_i$.

As a result, each gene across these four datasets have the same expression level, while

the variance is getting larger as $p_i$ getting larger. Since every two expression profiles were combined to make hybrids within each dataset, there would be no true *cis*-variation. The variance between alleles or among hybrid samples would be low in "`simu_null:1`" and high in "`simu_null:3`".

### 3.3.8 Bootstrap *cis* variation estimation

To test the discovery rate or the false positive rate with different replication number, we randomly choose (without replacement) $Nr$ replicates from all $N$ hybrids. For each level of replication (i.e., $Nr$), we did the resampling from these $N$ hybrids for $t$ times. Each time, we calculated $e_{cis}$ and its 95% confidence interval using maximum likelihood method (see §3.3.6). If a gene's $\log_2(e_{cis})$ confidence interval overlap with 0 ($e_{cis} = 1$), we classify it as a significant *cis*-variant gene. Table 3.1 shows the $N$ (number of hybrid samples), $Nr$ (Number of replicates tested), and $t$ (number of samplings) for each dataset.

### 3.3.9 Data availability

The allele-specific expression data and the short/long DNA reads to assemble the genomes are available at NCBI with the BioProject number PRJNA554649. The scripts mentioned and the summarized read count tables can be found at `https://github.com/xinwenz/yeastAse_bioinfo/tree/master/scripts_readcountsTable`.

## 3.4 Results

### 3.4.1 Assembly of reference genomes

Read mapping biases related to using only a single reference genome will lead to biases in allele-specific expression inference [13]. To mitigate such bias, we constructed two reference quality de novo genome assemblies of the parental strains used in this study, YPS128, and RM11-1a.

The contiguity, completeness, and accuracy of our assemblies are quite high (Table 3.2 and Figure 3.1). Both assemblies exhibit a high level of contiguity, with the majority of chromosomes being covered by one or two contigs, comparable to that of the *Saccharomyces cerevisiae* S288C Reference R64-1-1 (Table 3.2 and Figure 3.1). The BUSCO score assesses genome assembly completeness by identifying conserved single copy orthologs [63]. Both assemblies compare favorably to the yeast community reference genome (Table 3.2: RM11-1a: 93.7%; YPS128: 94.4%; R64: 93.9%). The QV scores we calculate reflect the basepair-level concordance between an assembly and Illumina short reads [28]. While the new assemblies are both quite accurate, due to the lower coverage and noisier long reads used in assembling RM11-1a, its assembly exhibited a lower QV even after polishing (Table 3.2: RM11-1a: 35.6; YPS128: 60.0).

### 3.4.2 Allele-specific RNAseq

We sequenced 20 replicates of hybrid mRNA samples. The 20 samples were used independently to construct 20 barcoded libraries that were pooled into a single sequencing experiment. After demultiplexing, we obtained 1.531 billion 75-bp paired-end reads. We then counted the allele-specific counts for each gene using the SNPs/Indels between YPS128 and

RM11-1a genomes. Mapping bias was eliminated by using both YPS128 and RM11-1a genomes as references in the mapping step and filtering out suspect SNPs/indels (Figure 3.10). We then discarded two replicates exhibiting the lowest correlation with other replicates (Figure 3.11), and finally obtained 18 replicates of allele-specific gene read counts for 4,710 genes.

### 3.4.3 The beta-binomial distribution models *cis*-expression sampling variation better than the binomial distribution

To assess the performance of two models, we additionally simulated 6 hybrid null datasets lacking true *cis*-variation (for details, see §3.3.7 & Table 3.1). For each dataset (Table 3.1), we applied our inference machinery to estimate the *cis*-variation parameter and its 95% confidence interval for each gene. As the null data exhibits no true *cis*-variation, any significant expression should be caused by false positives. We then plot the rate of rejecting null hypothesis (which reduced to the false positive rate in the null simulations) against replication to examine the behavior of the models as power increases (Figures 3.2, 3.3, 3.4). The beta-binomial model exhibited a false positive rate closer to the prediction than the binomial model in null datasets. However, for both models, the performance was poor when there was little replication (Figures 3.2, 3.3).

#### Inference on a highly replicated dataset without genetic or environmental variation

One validation of our model makes use of data from a yeast expression experiment comprising 44 biological replicates of a single haploid yeast strain under the same condition [18]. Pairs of expression profiles were combined into synthetic/in silico hybrid samples by permuting pair assignments such that the two alleles within one synthetic hybrid do not have any true

*cis* variation while retaining the sample variation between them (for details, see §3.3.7 & Table 3.1). This hybrid dataset was labeled "`Gier2015`", yielding 1,892 permuted synthetic hybrid samples ($44 \times 43 = 1,892$).

To test whether increasing replication improves *cis* estimation, we randomly sampled $Nr$ replicates without replacement from the 1,892 synthetic hybrids, performed *cis* parameter inference, and calculated the false positive rate. $Nr$ ranged from 1 to 35 for this dataset. For each level of replication (i.e., $Nr$), we sampled, as described above, 150 times to determine the distribution of the false positive rate (Figure 3.2a; see §3.3.8).

We generated and analyzed another hybrid dataset "`Xinw2018_yps`" following a similar approach to "`Gier2015`", but using our own expression experiment. The 18 expression profiles of the YPS128 allele were extracted from the 18 hybrid samples (2 of the 20 replicates are removed due to being outliers as measured in terms of exhibiting low correlation with other replicates), and pairs of profiles were combined into 306 ($18 \times 17 = 306$) synthetic hybrids (Figure 3.2b).

Our results demonstrate that the binomial model consistently rejects the null hypothesis at an elevated rate for $\alpha = 0.05$, exhibiting a consistent rejection rate across levels of replication (Figures 3.2a–3.2b). The beta-binomial model consistently exhibits a rejection rate that is lower than that of the binomial model. However, the beta-binomial does show some variation in rejection rate at low replication. In particular, for low replication in both the "`Gier2015`" and "`Xinw2018_yps`" datasets, the beta-binomial model shows an excess rate of rejection that subsides as replication increases.

The severity in underestimating variance using the binomial model depends on the underlying variance among replicates. The rejection rate of the binomial model can vary from 20% (Figure 3.2a) to as high as 30% (Figure 3.2b). Increased replication seems to have little effect on diminishing this problem. In contrast, the false-positive rate in the beta-binomial

model improves as replication increases. The increasing of the false-positive rate at the beginning of Figure 3.2a is likely an artifact resulting from the starting point of the maximum likelihood optimizer (see §3.3.6). Other than this artifact, the beta-binomial model also appears to underestimate the variation among replicates with fewer replicates, leading to high false-positive rate, though reduced as compared to the binomial model. The rejection rates improve with sufficient replication, asymptoting towards the significant level $\alpha$.

**De novo null simulation**

Although the null datasets we generated by randomly pairing real experimental replicates exhibit no true *cis* variation, there is the potential for unknown confounding factors that were not controlled. We therefore simulated four hybrid datasets for 5,000 genes from the gamma-Poisson distribution ("`simu_null:1-4`"), with the same expression level between alleles and explicit overdispersion parameters so that we can study the behavior of overdispersed expression data in the absence of differential gene expression.

The gamma-Poisson distribution (also known as the negative-binomial distribution) is widely used to model the read counts distribution among replicates [53, 54]. This distribution can be viewed as a Poisson distribution where the Poisson parameter is gamma distributed.

We simulated the expression profile of 5,000 genes across a wide number of expression levels under this model. The four different datasets with different over-dispersion profiles were generated by systematically varying the "$p$" parameter in the gamma-Poisson distribution for each dataset. We ensured that each gene maintained the same expression level across all four datasets. When p approaches zero, the Gamma-Poisson model approaches the Poisson model. When $p$ approaches one, the Gamma-Poisson model is strongly over-dispersed (for details, see §3.3.7: Null datasets from random number). We randomly paired samples within each of the four datasets following the same approach described above for "`Gier2015`"

and "`Xinw2018_yps`". This permitted us to vary the level of overdispersion and study the consequences for inference.

We set the p parameter to 0.1, 0.4, and 0.8 for the first three datasets (`simu_null:1-3` respectively). As a result, the first dataset (`simu_null:1`) has the lowest over-dispersion with expression profiles (the closest to the Poisson model) whereas the third (`simu_null:3` ↪ ) is the most over-dispersed. For the final simulation (`simu_null:4`) we chose a uniform distribution of p parameters with a mean of 0.4 for the 5,000 genes to simulate the impact for genome-wide inference when a dataset has genes with different levels of overdispersion.

The false-positive rate for these four datasets shows a similar pattern as in "`Gier2015`" and "`Xinw2018_yps`". The binomial model shows an elevated false-positive rate that is not mitigated with increased replication (Figure 3.3). The degree of excess false positives is related to the simulated over-dispersion of each dataset. The binomial model has a ∼7% false-positive rate in "`simu_null:1`" which is only 2% higher than the expected 5% (Figure 3.3a), but it can be as high as 38% in "`simu_null:3`" (Figure 3.3c). The performance of the binomial model on a changing "$p$" (Figure 3.3d, fp ∼16%) is similar to the constant "$p$" with the corresponding mean with an ∼3% higher rate of false-positives (cf. Figure 3.3b, fp ∼13%).

With few replicates and low overdispersion, the beta-binomial demonstrates a lower false-positive rate than expected (Figure 3.3a), suggesting that it is overestimating the variance when only a few replicates are used. The reverse is true under the high overdispersion simulation, suggesting it is underestimating the variance (Figure 3.3c). However, the model consistently approaches $\alpha$ with increasing replication. This is likely because the overdispersion parameter $\theta$ is poorly estimated with only a few replicates and relies on the initial arbitrary value in maximum likelihood optimizing, a situation that improves with higher replication.

**The effects of replication on ASE confidence intervals**

We then applied the $e_{cis}$ inference machinery on the experimental dataset of our 18 replicated hybrid samples (`Xinw2018`). More significant genes are discovered as more replicates are used. When all 18 replicates are used, we observe the rate of rejection appears to asymptote to ∼70% with the beta-binomial model. The number of significant genes from the binomial model exceeds the beta-binomial by ∼5% (Figure 3.4).

To explore the effect of gene expression level and number of replicates on the power, we chose 100 typical genes from each of the following categories: "lowly expressed genes" (average counts: 50–200); "intermediate expressed genes" (average counts: 400–600); and "highly express genes" (average counts: 1,500–3,500). We plotted the confidence intervals for each gene using the estimation calculated from four levels of replication (3, 6, 12, 18). Genes are ranked by their *cis* effect (Figure 3.5).

As expected, the beta-binomial model yields a wider confidence interval than the binomial model, reducing the false-positive rate. We also see that, as expected, when replication increases or with higher expression level, the confidence intervals narrow for both models, increasing the power (Figure 3.5).

### 3.4.4 *Cis* variation between YPS128 and RM11-1a strain is ubiquitous and often small in magnitude

The rate of rejection appears to asymptote to ∼70% (Figure 3.4) with beta-binomial model in the "`Xinw2018`" dataset, suggesting that ∼70% of the 4,710 genes we studied show evidence for expression variation, a marked increase compared to previous observations [14, 58, 45, 68, 2].

We then summarized the $e_{cis}$ distribution calculated form all 18 hybrid replicates with the beta-binomial model (Figure 3.6a, Table 3.3). The symmetry of the distribution of $\log_2(e_{cis})$ indicates that there are similar amount of genes affected by *cis*-regulatory variation in both directions. Approximately 70% of the genes (3,308 out of 4,710 ) exhibit significant *cis* variation ($|\log_2(e_{cis})| > 0$; $p < 0.05$). Notably, the *cis* effect in most of these significant genes is small in magnitude. Of the differentially expressed genes (Figure 3.6b), 70% exhibit *cis* variation in the range $0 < |\log_2(e_{cis})| < 0.2$, or less than a 1.15-fold difference. The genes with the *cis* variation $|\log_2(e_{cis})| > 1$ (i.e., a 2 fold difference) only comprise 3% of all significant genes.

## 3.5 Discussion

Inference of allele-specific expression differences from F1 hybrids is a widely used perspective to explore the evolution of gene expression. Many results have been reported for a wide range of individuals, populations, or species [68, 74, 43, 14]. Such inferences have been applied to questions about compensation between cis and trans variation [57, 39], stabilizing selection for expression level [20], and *cis*-effect in inter-specific/intra-specific expression variation [45, 52] and all depend in a central way on accurate measurement of *cis* variation. However, naive statistical models [53] and the tendency to misuse replication has limited the utility of allele-specific-expression inference.

In this work, we describe a beta-binomial model for estimation of *cis* expression variation in allele-specific studies. It is based on a more suitable gamma-Poisson distribution of read counts among replicated experiments and is capable of accommodating over-dispersion of expression. We demonstrate the advantage of the beta-binomial model over the binomial model with both experimental and simulated data. The results showed that, with sufficient replication, the beta-binomial model attains the nominal false positive rate while the bi-

nomial model consistently underestimates the variance leading to an elevated false-positive rate.

While, unlike the Poisson model, the gamma Poisson model permits the variance and mean to be independent, rigorous inference using the beta-binomial model derived from it still requires each allele to exhibit approximately the same variance-to-mean ratio (see §3.3.6: Beta-binomial Model). This limitation can be addressed by assigning different over-dispersion parameters for each allele, but inference becomes more complex. In any event, the good performance of the beta-binomial model suggests that potential improvement for $e_{cis}$ estimation is limited.

The trade-off between the false-positive rate and power still holds in these two models. We used the significant gene list from our best estimates (i.e., the beta-binomial model with all 18 replicates) as a gold standard to explore the relative power of both models (Figure 3.12). The binomial model has higher power than the beta-binomial model in all levels of replication. Of course, even the best statistical model would by definition exhibit $\alpha \times 100\%$ false positives. If we assume the 18 replicate beta-binomial model has 100% of power (Figure 3.12), then the proportion of true negatives that yields a false positive rate of 0.05 is $(1-0.702)/(1-0.05) = 0.314$. The 18 replicate binomial model rejects the null hypothesis 74.5% of the time, implying its false positive rate is 18% $(1-(1-0.745)/0.314 = 0.18$, assuming 100% power), which is consistent with our simulations (Figure 3.2a).

We uncovered many more *cis*-affected genes than previous intra-specific studies of yeast, where the proportion varies between 6%–29% [14, 58, 45]. The main culprit is likely lower power in previous studies, although we also used YPS128 rather than the BY4741 strain common in previous studies. Figures 3.4, 3.5, 3.12 demonstrate that adding more replicates increases the power and the relative difference in the discovery rate can be as high as 55% (Figure 3.4). Results from previous studies using one or two replicates yield comparable numbers of genes differentially expressed in *cis* (Figure 3.4, the left-most two points of the

Binomial model), suggesting that the difference in our results is of higher power to detect smaller magnitude changes.

Our results quantify the advantage of the beta-binomial model over the binomial model in detecting *cis* variation. The beta-binomial model estimates variance accurately and also has high statistical power as long as sufficient replicates are provided. Thus, our high replicate experiment describes an accurate and complete landscape of *cis* variation between YPS128 and RM11-1a. We recommend a beta-binomial model should for use in future allele-specific experiments and predict it will reveal an abundance of *cis* variation that previously remained hidden.

## 3.6 Acknowledgement

We thank A. Long, and TY. Wang for providing the yeast strain, T.Tsuboi for assistance in making yeast hybrid and culturing, J.Shan, R.Linder, J.Baldwin-Brown for assistance in making DNA and RNA libraries, M.Chakraborty for suggestions in assembling the genomes, J.Schraiber, A.Ramaiah and N.Zhao for thoughtful comments on the manuscript.

The work was supported by US National Institutes of Health (NIH) grant R01GM123303-1 (J.J.E.), University of California, Irvine setup funds (J.J.E). This work was made possible, in part, through access to the High Performance Computing Cluster of University of California, Irvine.

| **Synthetic NULL dataset** | **Number of genes in the dataset** | **Number of haploid expression profile** | **Number of permutate hybrid samples ($N$)** | **Number of replicates tested ($Nr$)** | **Number of samplings for each $Nr$ ($t$)** |
|---|---|---|---|---|---|
| `Gier2015` | 6,023 | 44 | 1,892 | 1–35 | 500 |
| `Xinw2018_yps` | 4,710 | 18 | 306 | 1–20 | 150 |
| `simu_null:1` | 5,000 | 100 | 9,900 | 1–25 | 150 |
| `simu_null:2` | 5,000 | 100 | 9,900 | 1–25 | 150 |
| `simu_null:3` | 5,000 | 100 | 9,900 | 1–25 | 150 |
| `simu_null:4` | 5,000 | 100 | 9,900 | 1–25 | 150 |
| **Experimental dataset** | | | **Number of available hybrid samples ($N$)** | | |
| `Xinw2018` | 4,710 | - | 18 | 1–18 | 150 |

Table 3.1: Summary of the datasets used for this study (for details see §3.3.7). `Gier2015` and `Xinwe2018_yps` are null datasets simulated from replicate expression profiles. `Simu_null` ↪ `:1-4` are null datasets simulated from random number generator. `Xinw2018` are real experimental data.

The number of genes and number of replicated expression profiles (except `Xinw2018`, hybrid samples do not have haploid expression profile) are in column 2 & 3. The permuted hybrid samples are listed in column 4. The number of replicates are listed in column 5. The ranges were chosen somewhat arbitrarily, but were enough to see the trend. For level of replication, we did the resampling $t$ times shown in column 6.

| | **S.cer R64-1-1** | **Rm11-1a** | **Yps128** |
|---|---|---|---|
| Assembly size (Mb) | 12.16 | 11.95 | 12.09 |
| Number of contigs/scaffolds | 17 | 19 | 29 |
| Contig N50 (Mb) | 0.92 | 0.92 | 0.81 |
| Contig L50 | 6 | 6 | 6 |
| Contig N90 (Mb) | 0.44 | 0.43 | 0.44 |
| Contig L90 | 13 | 13 | 14 |
| Busco score | 93.9% | 93.7% | 94.4% |
| Complete Busco | 1,351 | 1,347 | 1,358 |
| Fragmented Busco | 38 | 37 | 33 |
| Missing Busco | 49 | 54 | 47 |
| QV score | - | 35.6 | 60.0 |

Table 3.2: The contiguity, completeness, and accuracy of YPS128 and RM11-1a genomes.

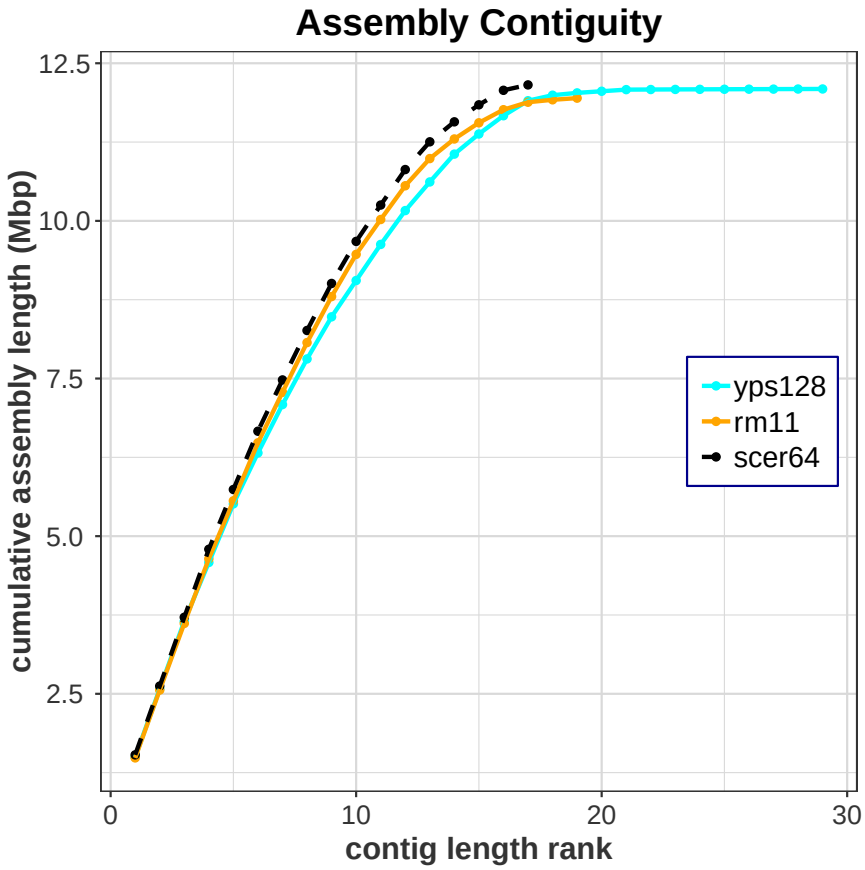
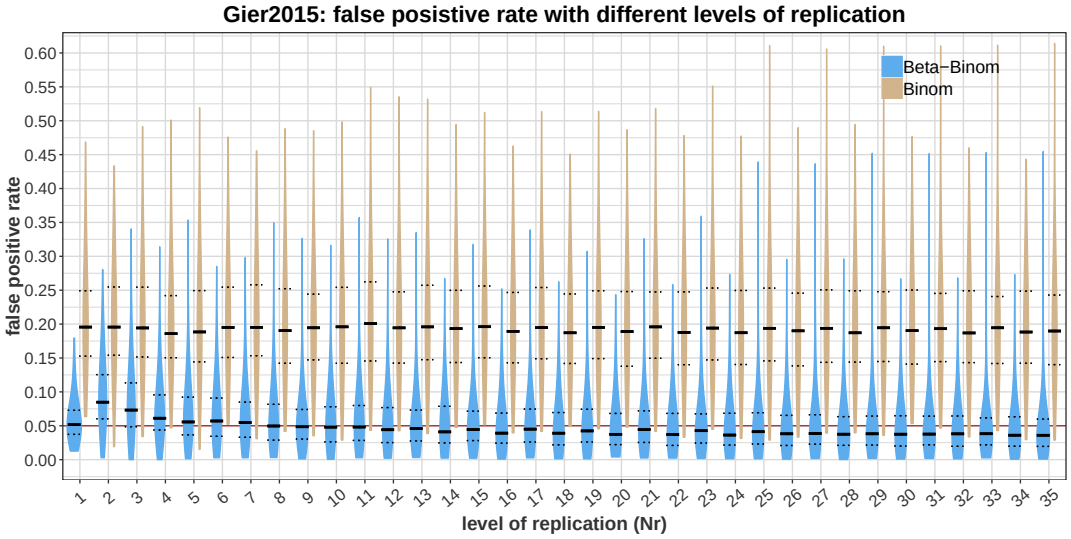
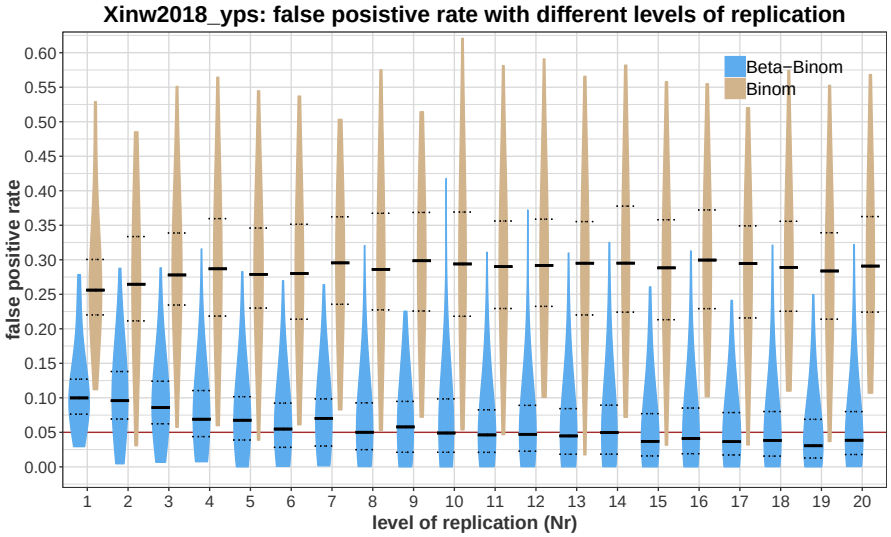



Figure 3.1: The contiguity of our assemblies is comparable to that of the *Saccharomyces cerevisiae* S288C Reference R64-1-1. Contigs are ranked from longest to shortest. Their cumulative sum of length are shown on Y axis in mega bases.




(a)




(b)

Figure 3.2: False positive rate with different number of replicates. The blue and brown violin plot in each level of replication show the distribution of false positive rates from $t$ (see Table 3.1) sampling results. The red horizontal line is the expected false positive rate of 0.05 ($\alpha = 0.05$). The solid and dot lines on each plot are the median, 25% quantile and 75% quantile.

(a) The binomial model consistently rejects the null hypothesis at a rate around 20%. The beta-binomial model consistently exhibits a rejection rate that is lower than that of the binomial model and is getting closer to the expected 5% as more replicates used. The rejection rates of beta-binomial model improve with sufficient replication, approaching the significant level $\alpha$.
(b) The binomial model consistently rejects the null hypothesis at a rate around 28%. Similar to panel A, the beta-binomial model consistently exhibits a lower rejection rate and is getting closer to the expected 5% as more replicates used.

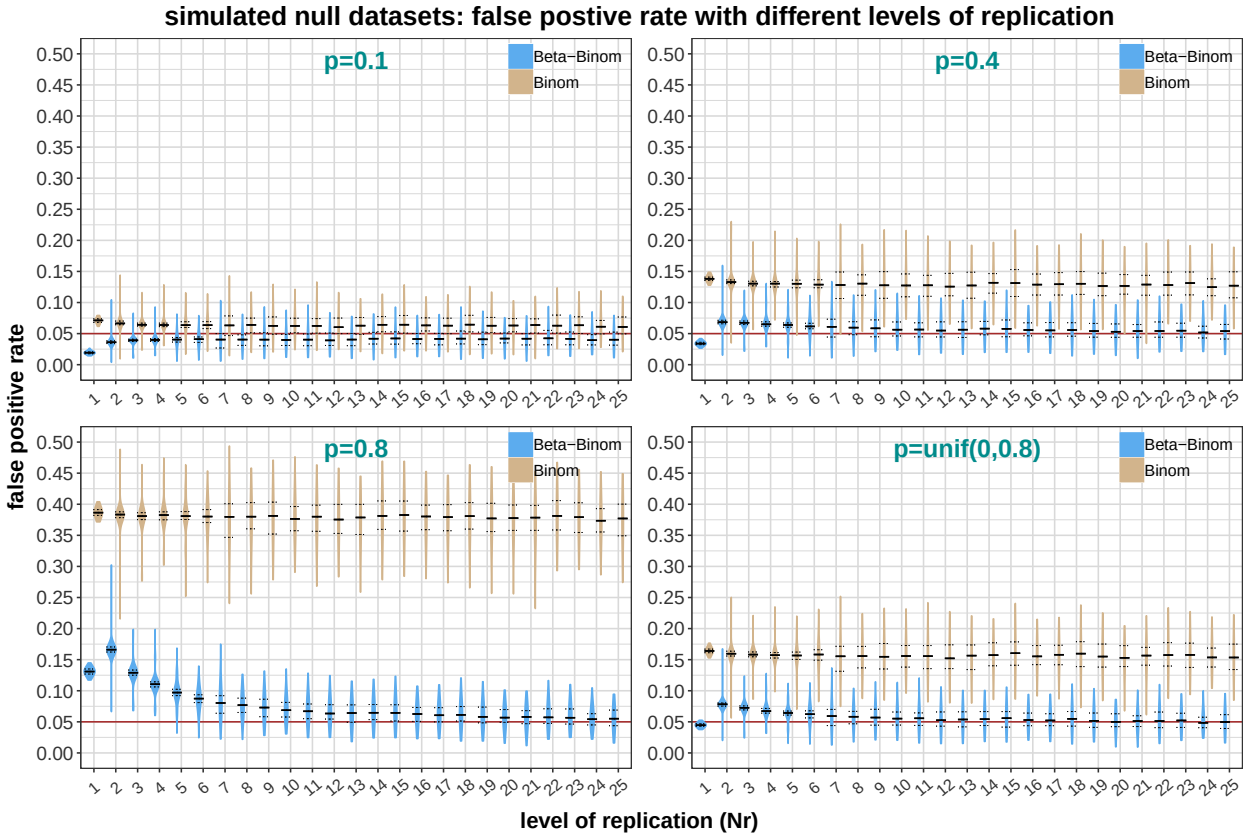



Figure 3.3: False positive rate with different number of replicates in `simu_null:1-4`. The degree of excess false positives is related to the simulated over-dispersion of each dataset ($p$ is a parameter controls the over-dispersion: when p approaches zero, the gamma-Poisson model approaches the Poisson model; when p approaches one, the gamma-Poisson model is strongly over-dispersed). The binomial model has a consistent ∼7% false-positive rate in "`simu_null:1`" which is only 2% higher than expected 5% (panel $p = 0.1$), but it can be as high as 38% in "`simu_null:3`" (panel $p = 0.8$). The performance of the binomial model on a changing "$p$" (panel $p \sim \text{unif}(0, 0.8)$) is similar to the constant "$p$" with the corresponding mean except a ∼3% more false-positives (panel $p = 0.4$). With few replicates and low overdispersion, the beta-binomial demonstrates a lower false-positive rate than expected (panel $p = 0.1$). The reverse is true under the high overdispersion simulation. The beta-binomial model consistently approaches $\alpha$ with increasing replicates.

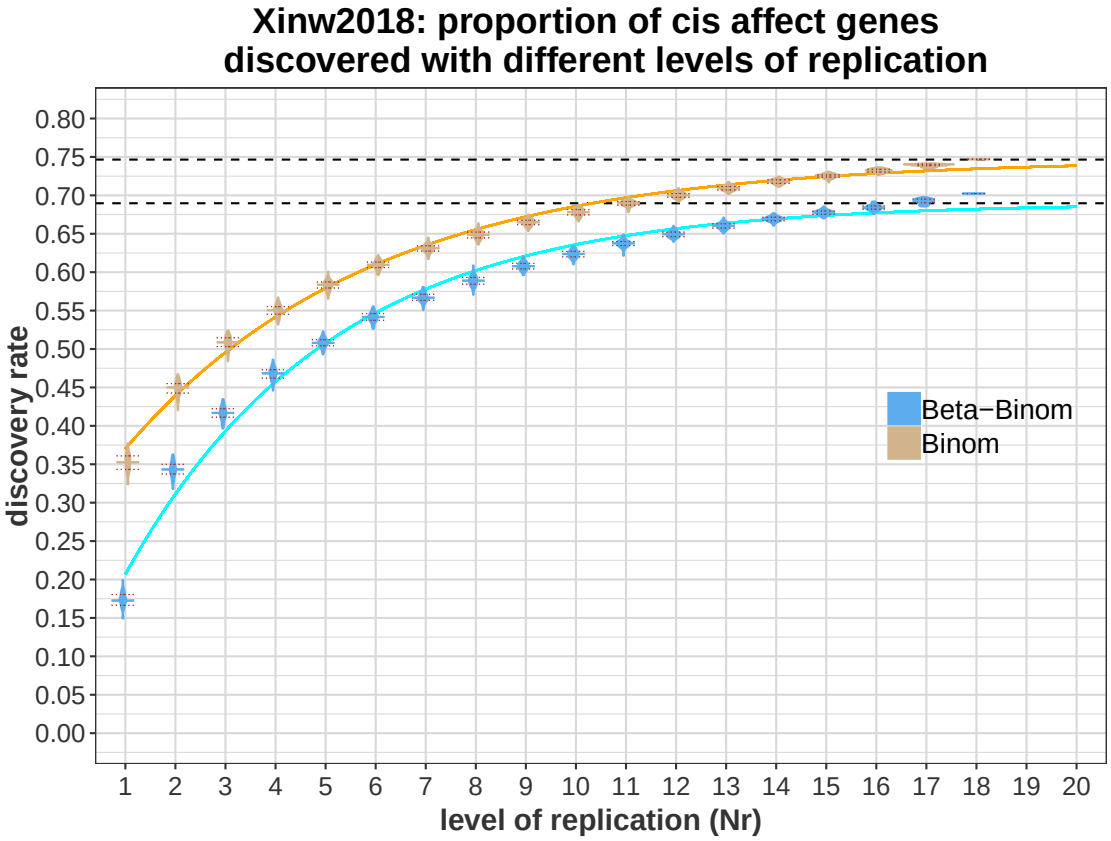



Figure 3.4: The discovery rate of the dataset "`Xinw2018`". More significant genes are discovered as more replicates used. When all 18 replicates are used, the rate of rejection appears to asymptote to ∼70% for the beta-binomial model and 75% for the binomial model. The two asymptotic lines were drawn by fitting the data to a negative exponential model.

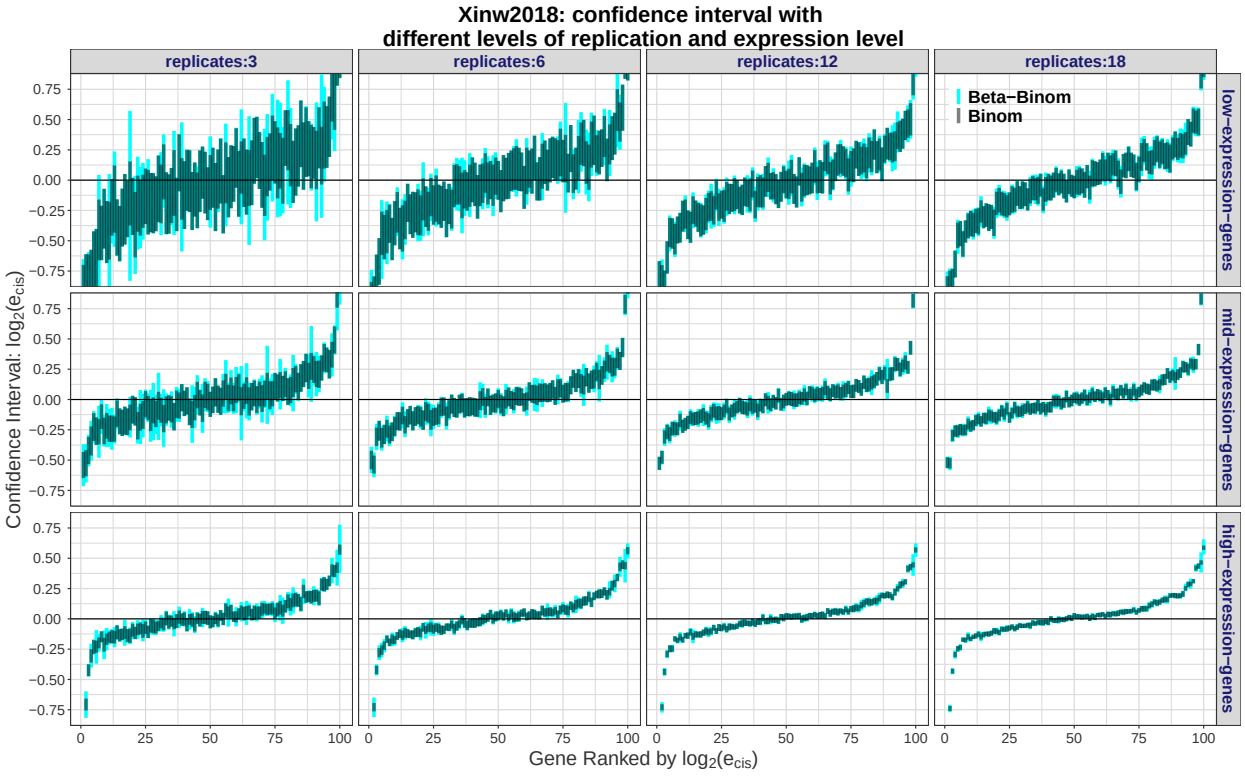
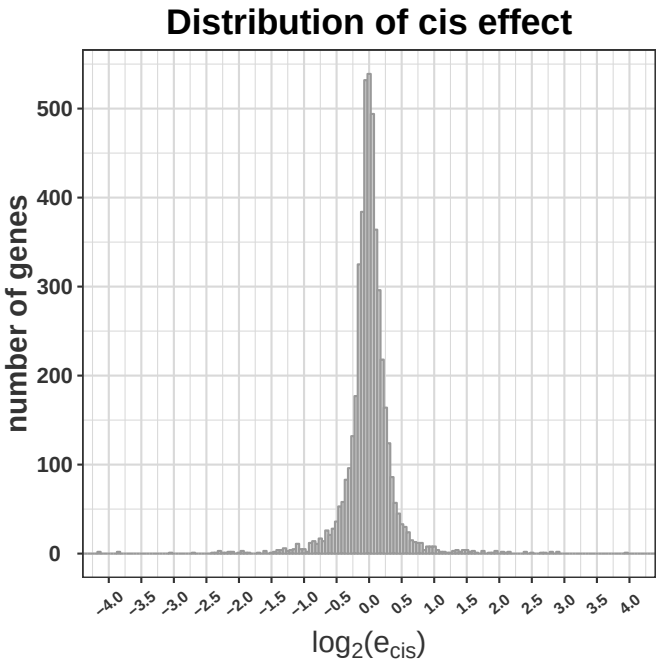
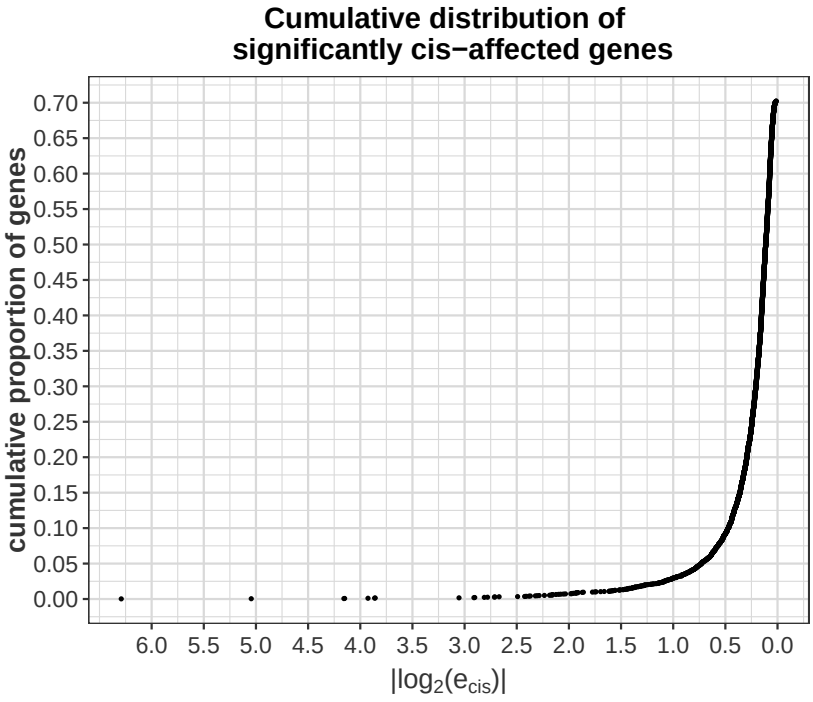



Figure 3.5: The effect of gene expression level and number of replicates on the inference power. 100 typical genes of low expression (average counts: 50–200), mid expression (average counts: 400–600) and high expression (average counts: 1,500–3,500) are plotted for their confidence intervals using the estimation calculated from four levels of replication (3, 6, 12, 18). The genes are ranked by their cis-effect: $\log_2(e_{cis})$. Grey is for binomial model; Blue is for beta-binomial model.




(a)




(b)

Figure 3.6: (a) The $\log_2(e_{cis})$ distribution from all 18 hybrid replicates with beta-binomial model.
(b) The cumulative proportion of significantly *cis*-affected genes. The *cis*-affected genes are sorted by their *cis* effect, from largest to smallest. The cumulative proportion shows that most significant genes have a *cis*-effect of small magnitude.

| **4,710 total genes** | Number of significant genes (significant level: 0.05) | Proportion of all genes |
|---|---|---|
| $\lvert \log_2(e_{cis}) \rvert > 0$ | 3,308 | 70.2% |
| $\lvert \log_2(e_{cis}) \rvert > 0.2$ ($\sim$1.15 fold change) | 1,008 | 21.4% |
| $\lvert \log_2(e_{cis}) \rvert > 0.5$ ($\sim$1.4 fold change) | 303 | 6.4% |
| $\lvert \log_2(e_{cis}) \rvert > 1$ (2 fold change) | 101 | 2.1% |

Table 3.3: The number of genes and their proportion with different *cis*-effect magnitude.

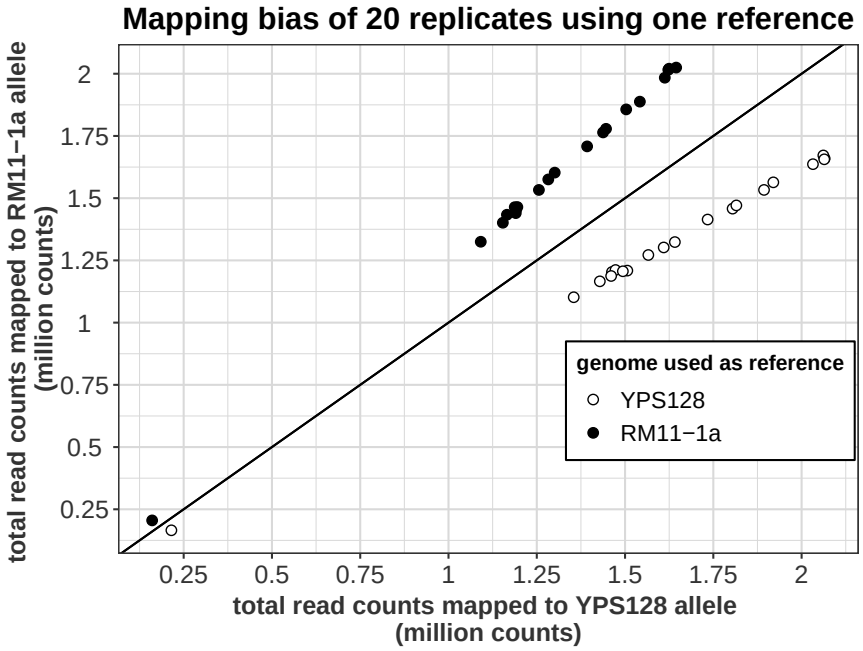



Figure 3.7: The mapping bias of DNA read counts when using only one assembly as the reference genome.

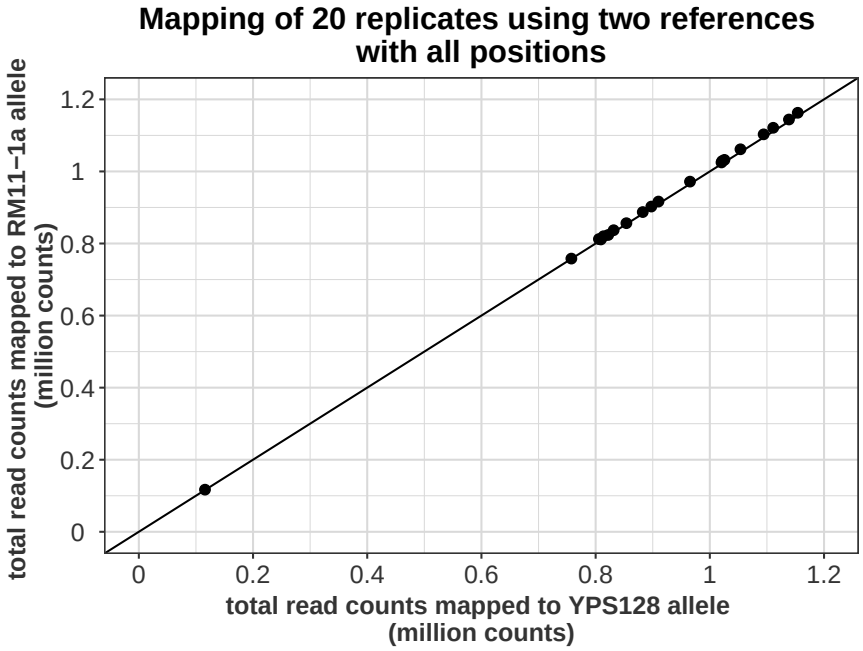



Figure 3.8: The mapping bias are mostly removed when using both assemblies as reference genomes.

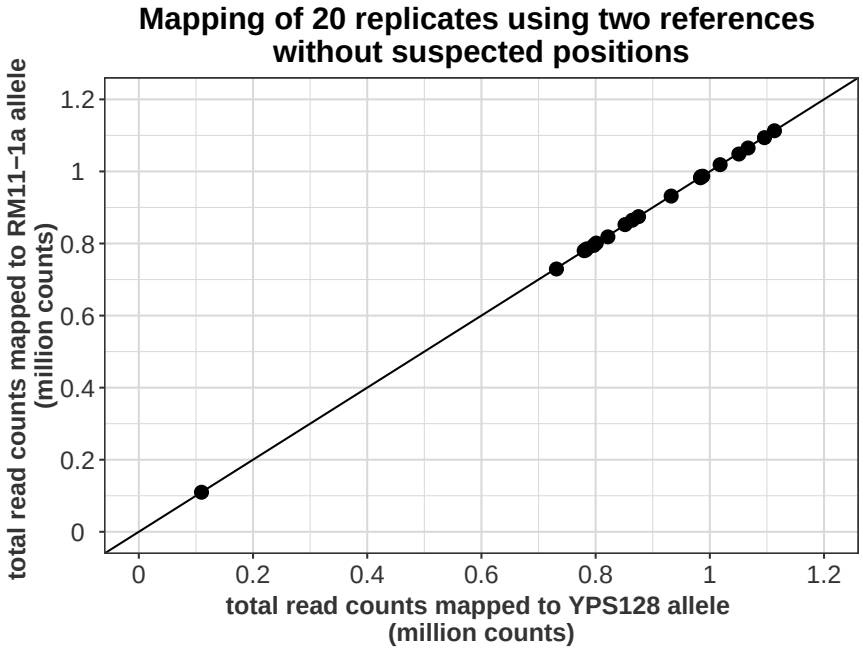



Figure 3.9: The mapping bias of DNA read counts are further eliminated by filtering out suspected variant positions.

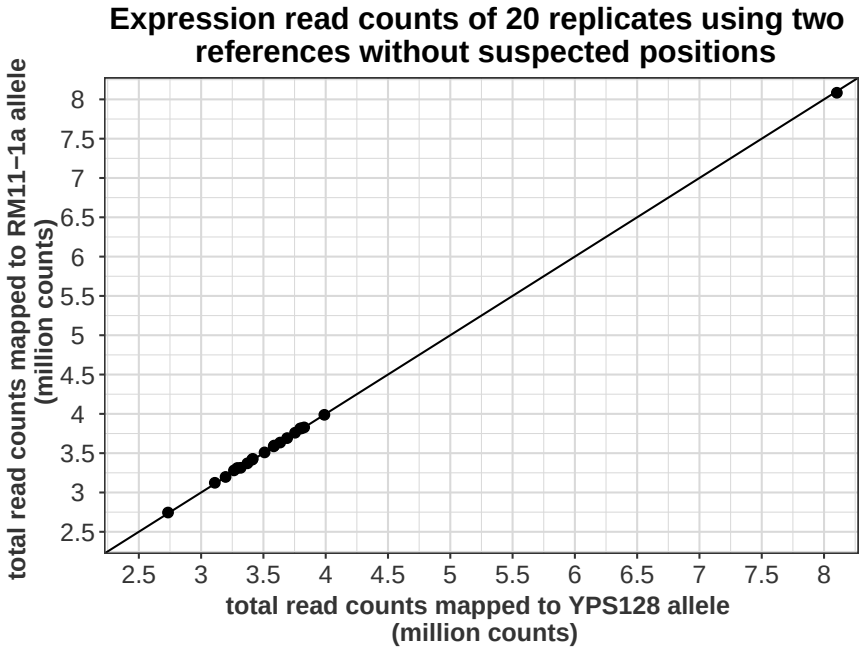



Figure 3.10: By using two references and removing all suspected variant positions, no mapping bias shown up for the RNA read counts.

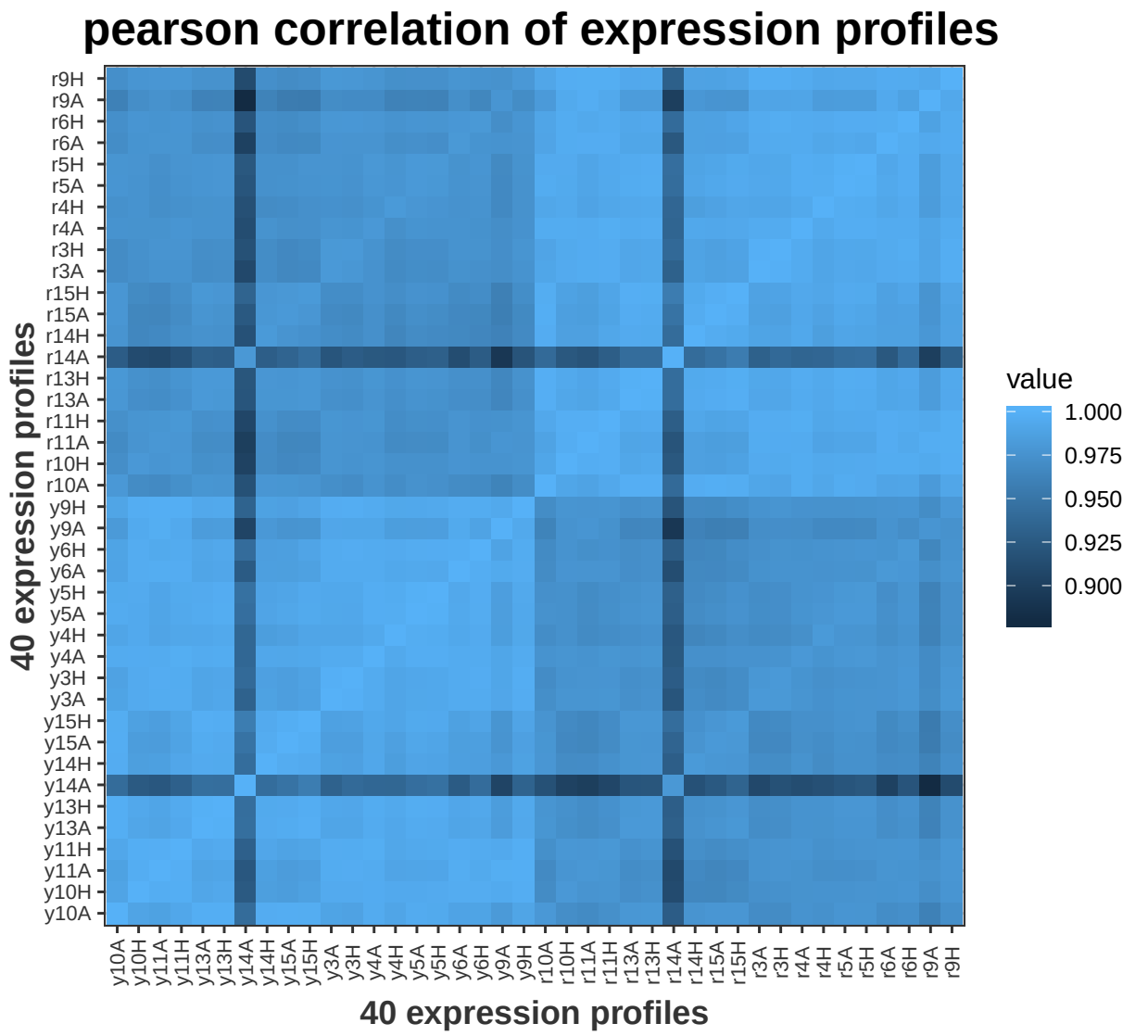



Figure 3.11: Correlation of 40 expression profiles. Profiles: r9A, y9A, r14A, and y14A are removed for downstream analysis because of the low correlation with other replicates.

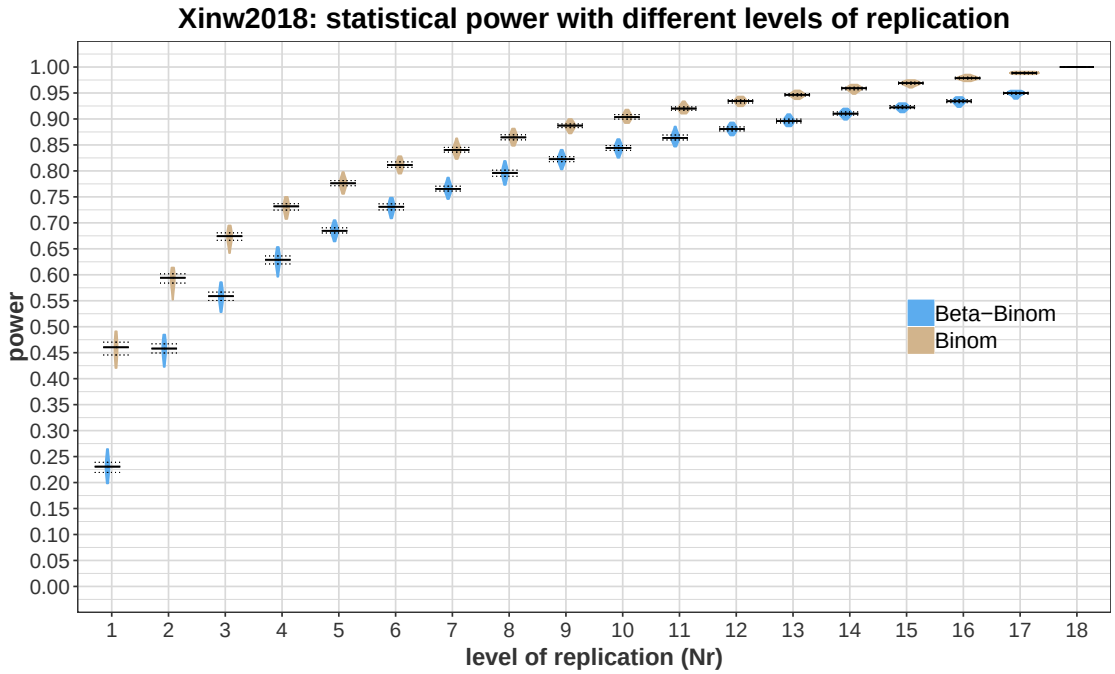



Figure 3.12: The statistical power of identifying significant *cis*-affected genes are plotted against increasing level of replication. The significant gene list from the best estimates (i.e., the beta-binomial model with all 18 replicates) was used as a gold standard. The power increases for both models as more replicates used. Binomial model has higher statistical power than beta-binomial model.

# Chapter 4

# Inferring Compensatory Evolution of *cis* and *trans* Regulatory Variation

This project discusses the commonly observed compensatory evolution in expression regulatory control. It shows with statistical principles that the observed compensatory evolution might just be a measurement artifact. It then discusses an improved method and demonstrates the reduction of the negative-correlation (an indicator of compensatory evolution) with both simulated and public data. Part of the work is published on *Trends in Genetics* under the title "Inferring compensatory evolution of *cis* and *trans* regulatory variation". A few materials was added to give out a comprehensive explanation of the correlation error and also to conform the format of the dissertation.

## 4.1 Abstract

The variation of gene regulation can be analyzed by decomposing the causing variants into *cis* variants and *trans* variants. We can estimate the *cis*/*trans* contribution (*cis*/*trans* effect)

using allele-specific expression experiments in hybrid and parental cells/individuals. Many previous studies observed a negative correlation between *cis* and *trans* effect and based on that, inferred that compensatory evolution is common for the gene regulatory evolution. However, the observed negative correlation can just be a measurement artifact. The *cis* and *trans* effect, which ideally should be estimated independently, have been estimated dependently in these results.

In this study, we first discuss the relationship between observed correlation and actual correlation using statistical principles, showing that the negative correlation is inherited in the previously used measurement scheme. We then discussed the new 'independent hybrid' scheme and demonstrated its advantage in reducing the correlated error with both simulated and public data.

## 4.2 Introduction

Genetic variation in gene regulation is an important source of phenotypic variation, contributing to human phenotypes and diseases [34, 33] as well as evolution within and between species [8, 15]. Expression variation between two individuals can be partitioned into diffusible *trans* elements (e.g., transcription factors) or non-diffusible *cis* elements (e.g., linked regulatory sequences like promoters or enhancers) [15]. By taking advantage of genetic crosses, we can gain insight into the mechanistic basis of expression variation that differentiates individuals [73, 74]. Because parental genotypes share a single cellular compartment in F1 hybrids, they also share all diffusible regulatory factors. Thus, expression variation between alleles in an F1 hybrid reflects the portion of variation between the parents due to *cis* factors alone. The remaining portion of variation between parents not explained by variation in the F1 hybrids is due to variation in *trans* factors. Conceptually, this leads to the mechanistic perspective that allele specific expression (ASE) variation in F1 hybrids is

equivalent to variation in *cis* elements whereas ASE variation in parents is a combination of variation in *cis*+*trans* factors [73]. By measuring the expression variation in both parents and their F1 hybrids, we can estimate the contribution of *cis* elements and *trans* factors to expression variation.

This ASE perspective facilitates estimation of important expression parameters on a genome scale [43, 14], providing abundant fodder for making mechanistic inferences on the genetic basis of expression variation within and between species. However, an article in this issue of Trends in Genetics points out that, when *cis* and *trans* estimates share common F1 hybrid samples, they will be negatively correlated via error shared from the hybrid data [16]. One important consequence of this observation is that spurious inferences of compensatory evolution between *cis* and *trans* factors will occur when correlated error is not accounted for. This is because this type of compensatory evolution is defined as a negative relationship between *cis* and *trans* variation. As [16] points out, many studies continue to make precisely this error regarding compensatory evolution, and consequently, a solution is urgently needed. [16] argues that the simplest solution to this problem is to estimate *cis* and *trans* parameters from independent replicates of hybrid data so that error is no longer correlated. Indeed, an ASE inference framework formulated by [14] recommends correcting for error in just this way (cf Figures 2 and S2 from [14]).

In this study, we first discuss the relationship between observed correlation and actual correlation using statistical principles, showing that the negative correlation is inherited in the previously used measurement scheme. We then discussed the new 'independent hybrid' scheme and demonstrated its advantage in reducing the correlated error with both simulated and public data.

## 4.3 Materials and Methods

### 4.3.1 Synthesis hybrid and parental samples with no true *cis* variance

We generated correlated and independent ASE datasets by partitioning 48 biological replicates of expression data from haploid yeast [18] (for details, see the flow chart of Figure 4.3). Four replicates were discarded because of the poor quality reported in the paper. To produce a dataset resulting in correlated estimates of *cis* and *trans* variation, the 44 remaining samples were partitioned into four subsets of 11 samples each, representing two "alleles" (strain 1 and strain 2) by two "conditions" (hybrid and parental). To produce a dataset resulting in independent estimates of *cis* and *trans*, 42 of the 44 remaining samples were partitioned into six subsets of 7 samples each, representing two "alleles" (strain 1 and strain 2) by three "conditions" (hybrid 1, hybrid 2, and parental).

### 4.3.2 Calculate *cis*/*trans* with both methods

Estimates of *cis*/*trans* were calculated on the sums of individual partitions according to [14], with the modification that dHybrid and dParent were calculated according to the expression:

$$d = \frac{\sum_{\text{genome}} \text{strain 1 reads}}{\sum_{\text{genome}} \text{strain 2 reads}}$$

### 4.3.3 Calculate *cis*/*trans* correlation of both methods

For a random split in the simulation process, Pearson correlation coefficient was calculated in R, and a hypothesis test on the significance of the correlation was also applied.

### 4.3.4 Correlation coefficient distribution of both methods

Since different partition of the 44 profiles produces different *cis*/*trans* calculation and thus affects their correlation coefficient, the distributions of correlation coefficient are more suitable for comparing the two schemes. We then repeated the above process 500 times (Figure 4.3) and obtained 500 correlation coefficient data point for each scheme.

### 4.3.5 Test on true data of hybrid and parental samples

We obtained correlated and independent ASE datasets from supplemental datasets 1 and 2 in [14]. These are allele-specific expression count data from the co-culture and hybrid of two Saccharomyces cerevisiae strains: BY4741 and RM11-1a. We calculated *cis*/*trans* correlation and did hypothesis testing as mentioned above.

## 4.4 Results

### 4.4.1 Statistical principles for the correlated error in two methods of measuring *cis*/*trans* effect

#### Standard method

Let assume there is no *cis*-*trans* interaction (the *cis* mutation between two alleles has a consistent effect in whatever genetic background).

The expression ratio of a gene in a hybrid strain/sample is used for *cis* effect estimation and can be expressed by the following equation (the ratio in hybrid and *cis* effect are all

described in log space, the same applied for all the following equations).

$$\text{hybrid} = \hat{c} = c + \epsilon_h$$

The real *cis* effect is represented by $c$. The measurement from the hybrid cannot be exactly accurate as the real *cis* effect, so we assign an error term $\epsilon_h$ and assume that $\epsilon_h$ is distributed as a Gaussian variable with mean equals zero. The expression ratio of the two parental strains/samples can be represented in a similar way.

$$\text{parent} = c + t + \epsilon_p$$

The real *trans* effect is represented by $t$, and $\epsilon_p$ is the error term from measuring parental ratio, which is independent of $\epsilon_h$. If the same hybrid strain/sample was used in *trans* estimation, then we can express the *trans* estimation in the following equation:

$$\begin{aligned} \hat{t} &= \text{parent} - \hat{c} \\ &= c + t + \epsilon_p - (c + \epsilon_h) \\ &= t + \epsilon_p - \epsilon_h \end{aligned}$$

Because the value of $c$, $t$, $\epsilon_h$ $\epsilon_p$, are all independent of each other and the expectation of $\epsilon_h$

$\epsilon_p$ equals to zero, we can express the covariance of *cis-trans* estimation in the following way:

$$
\begin{aligned}
& \mathrm{Cov}(\hat{c}, \hat{t}) \\
= \; & \mathrm{Cov}(c + \epsilon_h, t + \epsilon_p - \epsilon_h) \\
= \; & \mathbb{E}[(c + \epsilon_h) \times (t + \epsilon_p - \epsilon_h)] - \mathbb{E}[c + \epsilon_h] \times \mathbb{E}[t + \epsilon_p - \epsilon_h] \\
= \; & \mathbb{E}[c{\cdot}t + c{\cdot}\epsilon_p - c{\cdot}\epsilon_h + t{\cdot}\epsilon_h + \epsilon_h{\cdot}\epsilon_p - \epsilon_h^2] - \mathbb{E}[c]{\cdot}\,\mathbb{E}[t] \\
= \; & \mathbb{E}[c{\cdot}t] - \mathbb{E}[\epsilon_h^2] - \mathbb{E}[c]{\cdot}\,\mathbb{E}[t] \\
= \; & \mathrm{Cov}(c, t) - \mathrm{Var}(\epsilon_h)
\end{aligned}
$$

This shows that the *cis-trans* covariance observed can be different from the true *cis-trans* covariance qualitatively. When there is no correlation or a little bit positive correlation between true *cis* and *trans* effect, the observed correlation could still be negative.

**Independent hybrid method**

Same as before, assume there is no *cis-trans* interaction. We use one hybrid replicate for *cis* estimation and another hybrid replicate along with one parent replicate for *trans* estimation.

$$
\begin{aligned}
\mathrm{hybrid}_1 &= \hat{c}_1 = c + \epsilon_s + \epsilon_{h_1} \\
\mathrm{hybrid}_2 &= \hat{c}_2 = c + \epsilon_s + \epsilon_{h_2}
\end{aligned}
$$

In the above equation, $c$ is the real *cis* effect as in box1. If hybrid1 and hybrid2 are technique replicates, they may have a shared systematic error. We use $\epsilon_s$ to represent this systematic error and the remaining error are represented by $\epsilon_{h_1}$ and $\epsilon_{h_2}$. Let's assume that $\epsilon_s$ is also Gaussian distributed with mean equals zero. The equation for the parent sample is the same as box1.

The hybrid1 sample is used for *cis* estimation, and the hybrid2 sample was used in *trans*

estimation, so we can express the *trans* estimation in the following equation:

$$\begin{aligned}
&\hat{t} = \text{parent} - \hat{c}_2 \\
&= c + t + \epsilon_p - (c + \epsilon_s + \epsilon_{h_2}) \\
&= t + \epsilon_p - \epsilon_s - \epsilon_{h_2}
\end{aligned}$$

Because the value of $c$, $t$, $\epsilon_{h_1}$, $\epsilon_{h_2}$, $\epsilon_s$, are all independent of each other and the expectation of $\epsilon_{h_1}$, $\epsilon_{h_2}$, or $\epsilon_p$ equal to zero, we can express the covariance of *cis-trans* estimation in the following way:

$$\begin{aligned}
&\text{Cov}(\hat{c}_1, \hat{t}) \\
&= \text{Cov}(c + \epsilon_s + \epsilon_{h_1}, t + \epsilon_p - \epsilon_s - \epsilon_{h_2}) \\
&= \mathbb{E}[(c + \epsilon_s + \epsilon_{h_1}) \times (t + \epsilon_p - \epsilon_s - \epsilon_{h_2})] - \mathbb{E}[c + \epsilon_s + \epsilon_{h_1}] \times \mathbb{E}[t + \epsilon_p - \epsilon_s - \epsilon_{h_2}] \\
&= \mathbb{E}[c \cdot t] - \mathbb{E}[\epsilon_s^2] - \mathbb{E}[c] \cdot \mathbb{E}[t] \\
&= \text{Cov}(c, t) - \text{Var}(\epsilon_s)
\end{aligned}$$

If there is no systematic error, the *cis-trans* covariance observed is the same as the true *cis-trans* covariance.

### 4.4.2 The reduction of correlated error by the independent hybrid method

In order to demonstrate the utility of this approach, we investigate two ASE datasets. The first is an artificial dataset designed to be devoid of genetic variation in gene expression, and is constructed purely from biological replicates of the same strain from [18]. The second involves genetically distinct strains from [14] and therefore potentially exhibits compensatory variation in gene regulation (for details, see §4.3).

### Synthetic hybrid and parental samples

Figures 4.1a–4.1b illustrate the estimation of *cis* and *trans* expression parameters both with and without correcting for correlated error in a representative random partition of the ASE dataset constructed without genetic variation between the parents. The negative correlation in panel a is large in magnitude and highly significant ($r = -0.67$, $p < 0.0001$) while that of panel b is small but in a positive direction ($r = 0.14$, $p < 0.0001$). Overall, when full biological replication is employed for 500 times, correlation coefficients of the independent hybrid scheme cluster around 0 (Figure 4.1c).

### Experimental hybrid and parental samples

Figure 4.2 illustrates the estimation of *cis* and *trans* expression parameters of experimental data obtained from [14]. Panel a is estimated with standard scheme. Panel b is estimated with the independent hybrid scheme correcting for correlated error. The negative correlation in panel a is large in magnitude and highly significant ($r = -0.46$, $p < 0.0001$) while in panel b, there is no significant correlation ($r = -0.028$, $p = 0.08$).

## 4.5 Discussion

Given that the approach in [14] places *cis* and *trans* expression parameters in a likelihood testing framework, it can address questions of compensatory evolution on a gene-by-gene basis in a way purely correlative approaches cannot. For example, in [14], the overall correlation between *cis* and *trans* was near zero in the independent estimates, offering no evidence for compensatory evolution (Figure 4.2b, $p$-value $= 0.08$), compensating for a spurious conclusion of rampant compensatory evolution suggested by the correlated estimates (Figure 4.2a, $r = -0.46$, $p$-value $< 10^{-15}$).

However, by employing independent estimates of *cis* and *trans*, individual genes with evidence for differential expression can be identified. Of the 850 genes significant for *cis* and/or *trans* in the independent dataset of [14], 55% (466/850 with a 95% binomial confidence interval on the proportion 51%–58%) fall into the compensatory category at a significance threshold of 1%. Under a model of random expression variation, only 50% (425) are expected to fall in compensatory categories – quadrants II and IV – by chance (16 genes were excluded that have a *cis* estimate of 0 and cannot be classified as compensatory or reinforcing). Thus, while no evidence for a negative correlation between *cis* and *trans* is apparent at the genome level, the statistical evidence might support the action of compensatory evolution above the background expectation for at most a small number of genes ($\sim$41). Alternatively, because of the nature of replication in [14] (replicate cultures were pooled before library preparation and subsequent replicates came from the same library), the variation associated with library preparation was not controlled, perhaps explaining the remaining small magnitude of excess compensatory evolution observed in the study. Clearly, however, a substantial proportion of the signal of compensatory variation was caused by correlated error arising from sequencing, as the method of [14] reduced the correlation from $-0.46$ to $-0.028$.

This approach illustrates the utility of accounting for correlated error in a statistical inference framework. The ability to make inferences on individual genes is an important advantage in carefully measuring the extent of compensatory evolution. Indeed, any time estimates of *cis* and *trans* are considered jointly to make biological conclusions, correlated error should be considered, not just in cases of compensatory evolution. Modern datasets should be even better suited to addressing such questions, as lower sequencing costs allow us to achieve higher and higher replication, not only eliminating the correlated error problem, but also improving statistical power. Indeed, it would be irresponsible not to replicate parental and hybrid treatments in future ASE studies.

## 4.6 Acknowledgment

The work was supported by US National Institutes of Health (NIH) grant R01GM123303-1 (J.J.E.) and University of California, Irvine setup funds (J.J.E).

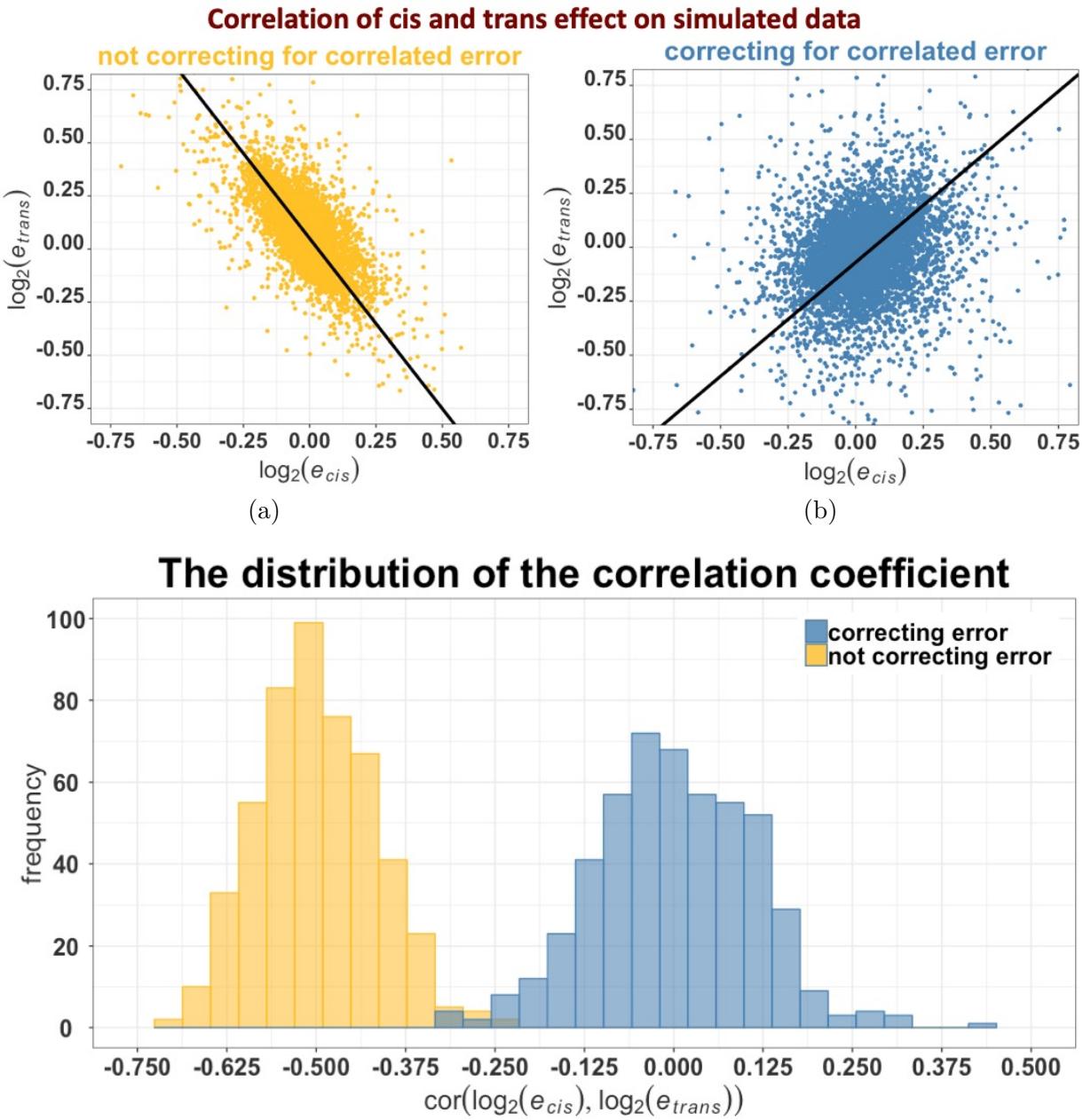
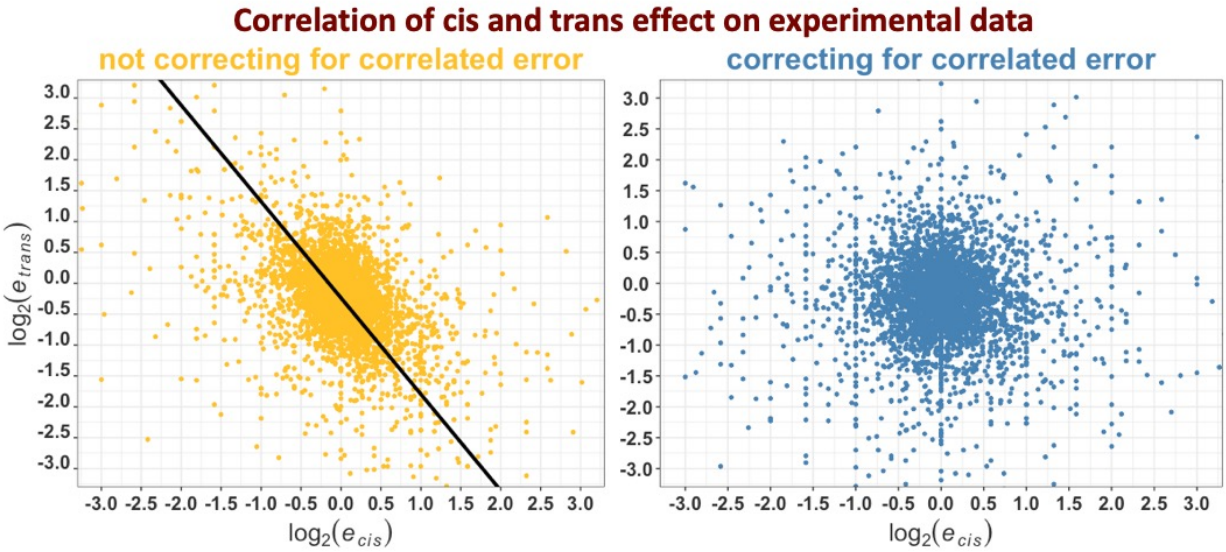



Figure 4.1: The effect of correlated error on estimation of *cis* and *trans* expression variation ratios. The data considered in the figure was compiled from partitions of a highly-replicated expression dataset in yeast [18] (for details see Materials and Method).

(a) Both *cis* and *trans* parameter estimates share a common sample of 11 hybrid individuals. (b) *Cis* parameters are estimated from one set of 7 hybrid individuals and *trans* parameters are estimated from a different set of 7 individuals. (c) Summary of $\boldsymbol{\tau}$ (Kendall rank correlation coefficient) for 500 randomly chosen partitions of both the correlated and independent estimation schemes. Panels (a) and (b) are representative instances of these random partitions.




(a) (b)

Figure 4.2: The estimation of *cis* and *trans* expression parameters of experimental data obtained from [14].

(a) Both *cis* and *trans* parameter estimates share a common hybrid sample. (b) *cis* parameters are estimated from one hybrid sample and *trans* parameters are estimated from a different hybrid sample.

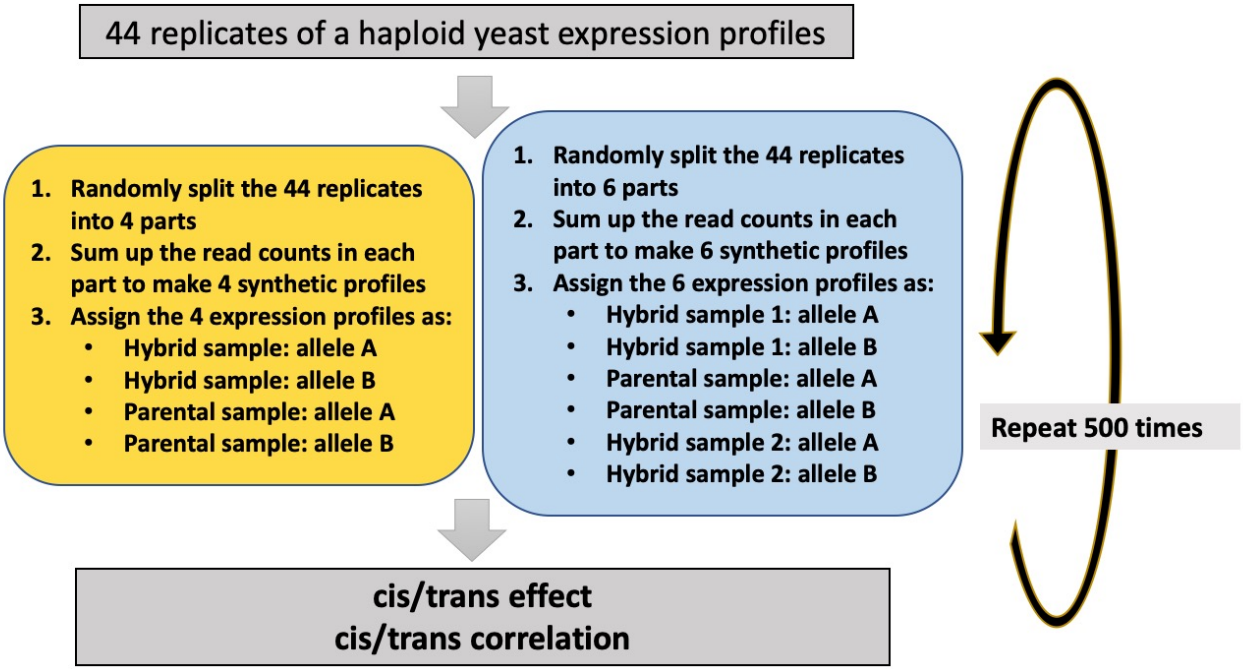



Figure 4.3: Work flow for the synthesis of hybrid and parental samples from 44 replications of haploid yeast expression. For the standard scheme (yellow), the 44 were partitioned into four subsets of 11 samples each, representing two "alleles" (strain A and strain B) by two "conditions" (hybrid and parental). For the independent hybrid scheme (blue), 42 of the 44 remaining samples were partitioned into six subsets of 7 samples each, representing two "alleles" (strain A and strain B) by three "conditions" (hybrid 1, hybrid 2, and parental). To get the distribution of correlation, 500 replicates was done.

# Chapter 5

# Conclusion

My three projects explore the *cis* and *trans* mechanisms for gene expression variation. The underlying cause of expression variation is denoted *cis* if it affects gene expression in a non-diffusible way, for example, the variation of the promoter sequences. On the other hand, the variation is defined as *trans* if it affects gene expression in a diffusible way, such as changes transcription factors [15, 62].

Gene duplication is an important mechanism in genome evolution. It directly provides new genetic material, which serves as the raw material for the origins of novel gene function [46]. When a new duplication occurs, it immediately changes the *cis*-element of the gene by doubling the RNA polymerase target. Some duplicates lose important elements as a consequence of the mutation itself. However, some duplicates preserve the complete sequence and are capable of expression. Gene expression is controlled by both *cis* and *trans* elements, so doubling the sequence does not mean the expression level doubles. There are several possible situations. One possible scenario is that gene has an important function and is tightly monitored by the regulatory network. As a result, the expression may be controlled by some feedback loop and the dosage-sharing model applies [56, 21]. Alternatively, the *trans*

effect is not independent of the *cis* effect [67, 62]. For example, the same regulatory network may have different effects on the two identical sequence copies, depending on the position of the sequence or the state of the chromatin in that part of the chromosome. Within the scope these expected confounders, we still hope to draw some general conclusions about the consequences for expression resulting newly duplicated genes.

In the paralog-specific expression project, I performed a preliminary survey of the expression level of 35 newly duplicated genes in A3 and A4 strains exhibiting complete duplication of their gene sequences (*Drosophila melanogaster*). We can not differentiate the RNA read counts for each paralog if they are 100% identical in the sequence. For these genes, we could only measure their total expression level. For a subset of genes, we can not only differentiate the reads counts but also tell whether the paralog is original or derivative. We observed that the two paralogs usually express differently. We also found that the new copies tend to express less than the original copy. The total expression does not show a clear relationship to their copy number, except that duplication typically leads to at least nominal increases in expression. However, the realized expression level may be higher or lower than that predicted by the copy number change.

Although these 35 genes are originally selected without regard to whether or not they contribute to fitness differences, they include a small group of duplicated genes that are candidates for positively selected or at least neutral. Since we selected mutations that are present one of A3 or A4 and not the other, we have imposed an ascertainment bias that makes them more likely to exhibit higher allele frequencies than mutations without such asertainment biases. As a consequence, these mutations are very likely to be enriched for beneficial alleles. This means that the expression change we observed sample biased towards containing beneficial expression changes and would not be suitable for predicting the effect of a duplication event. On the other hand, regardless of its affect on fitness, a mutation that changes a gene's expression is always of interest to molecular geneticists who try to connect DNA sequence to

the phenotype. In any event, it would be good to have population data instead of just two assemblies, so that we can measure the expression level across the full frequency spectrum and remove the ascertainment bias.

In the allele-specific expression project, I expand the paralog-specific project into a more general perspective of how to correctly measure *cis* variation. The number of biological replicates for the gene expression experiments has been discussed for many years. Many studies have been done to find out the balance between the accuracy of scientific results and economic efficiency [59]. For normal differential expression experiments, three to six replicates are acceptable. But for allele-specific experiments, the number of replicates and a model to estimate the variance between replicates has long been neglected. In principle, allele-specific expression experiment involves more precise measurement and more data filtering steps. Both require more replicates. In this project, I describe a new implementation of a beta-binomial model for the over-dispersed variance between replicates and demonstrate its usage in 20 replicates of allele-specific measurement on a yeast hybrid. With so many replicates, we can also down sample to demonstrate the consequences of using fewer replicates. This project only discusses the *cis* variation measurement, but in principle would also be applicable to *trans* variation measurement.

One important issue is the *cis*-*trans* interaction. The *trans* (regulatory network) variance is not completely independent of *cis* variance. The current measuring scheme assumes that the *cis*-*trans* effect is independent (one event of *cis* mutation always has the same effect on expression level in any *trans* background). The problem has been noticed by several studies [67], but the scheme is still popular for gene expression analysis. More sadly, *cis* and *trans* estimations are widely used for evolutionary inferences.

The project "Inferring compensatory evolution of *cis*/*trans* regulatory variation" in Chapter 4 tries to correct a measurement error in the inference process. It demonstrates that the previously used method introduces an artificial negative correlation between *cis* and *trans*,

which should not be used to infer compensatory evolution. However, even if this problem is fixed for future studies, the larger issue still exists. A clear method to analyze the *cis*/*trans* and the interaction between them and using quantitative instead of a verbal model for their evolutionary inference are very much in need.

# Bibliography


[1] S. Anders, P. T. Pyl, and W. Huber. HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, Jan. 2015.

[2] C. G. Artieri and H. B. Fraser. Evolution at two levels of gene expression in yeast. *Genome Res.*, 24(3):411–421, Mar. 2014.

[3] G. D. M. Bell, N. C. Kane, L. H. Rieseberg, and K. L. Adams. RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biol. Evol.*, 5(7):1309–1323, 2013.

[4] D. A. Bitton, M. J. Okoniewski, Y. Connolly, and C. J. Miller. Exon level integration of proteomics and microarray data. *BMC Bioinformatics*, 9:118, Feb. 2008.

[5] R. B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 102(5):1572–1577, Feb. 2005.

[6] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, Apr. 2002.

[7] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, Dec. 2009.

[8] S. B. Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, July 2008.

[9] M. Chakraborty, J. G. Baldwin-Brown, A. D. Long, and J. J. Emerson. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.*, 44(19):e147, Nov. 2016.

[10] M. Chakraborty, J. J. Emerson, S. J. Macdonald, and A. D. Long. Structural variants exhibit allelic heterogeneity and shape variation in complex traits. Sept. 2018.

[11] M. Chakraborty, N. W. VanKuren, R. Zhao, X. Zhang, S. Kalsow, and J. J. Emerson. Hidden genetic variation shapes the structure of functional elements in drosophila. *Nat. Genet.*, 50(1):20–25, Jan. 2018.

[12] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biol.*, 17:13, Jan. 2016.

[13] J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, Dec. 2009.

[14] J. J. Emerson, L.-C. Hsieh, H.-M. Sung, T.-Y. Wang, C.-J. Huang, H. H.-S. Lu, M.-Y. J. Lu, S.-H. Wu, and W.-H. Li. Natural selection on cis and trans regulation in yeasts. *Genome Res.*, 20(6):826–836, June 2010.

[15] J. J. Emerson and W.-H. Li. The genetic basis of evolutionary change in gene expression levels. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 365(1552):2581–2590, Aug. 2010.

[16] H. B. Fraser. Improving estimates of compensatory cis-trans regulatory divergence. *Trends Genet.*, Sept. 2018.

[17] E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. July 2012.

[18] M. Gierliński, C. Cole, P. Schofield, N. J. Schurch, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, 31(22):3625–3630, Nov. 2015.

[19] M. E. Hoballah, T. Gübitz, J. Stuurman, L. Broger, M. Barone, T. Mandel, A. Dell'Olivo, M. Arnold, and C. Kuhlemeier. Single gene-mediated shift in pollinator attraction in petunia. *Plant Cell*, 19(3):779–790, Mar. 2007.

[20] A. Hodgins-Davis, D. P. Rice, and J. P. Townsend. Gene expression evolves under a House-of-Cards model of stabilizing selection. *Mol. Biol. Evol.*, 32(8):2130–2140, Aug. 2015.

[21] H. Innan and F. Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.*, 11(2):97–108, Feb. 2010.

[22] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, June 1961.

[23] D. Kim, B. Langmead, and S. L. Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, 12(4):357–360, Apr. 2015.

[24] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, Apr. 2013.

[25] E. G. King, C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen, K. W. Broman, A. D. Long, and S. J. Macdonald. Genetic dissection of a model complex trait using the drosophila synthetic population resource. *Genome Res.*, 22(8):1558–1566, Aug. 2012.

[26] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, Apr. 1975.

[27] F. A. Kondrashov, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. Selection in the evolution of gene duplications. *Genome Biol.*, 3(2):RESEARCH0008, Jan. 2002.

[28] S. Koren, A. Rhie, B. P. Walenz, A. T. Dilthey, D. M. Bickhart, S. B. Kingan, S. Hiendleder, J. L. Williams, T. P. L. Smith, and A. M. Phillippy. Complete assembly of parental haplotypes with trio binning. Feb. 2018.

[29] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, 27(5):722–736, May 2017.

[30] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biol.*, 5(2):R12, Jan. 2004.

[31] K.-K. Lam, K. LaButti, A. Khalak, and D. Tse. FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. *Bioinformatics*, 31(19):3207–3209, Oct. 2015.

[32] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, Mar. 2012.

[33] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. C. 't Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, Sept. 2013.

[34] T. I. Lee and R. A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, Mar. 2013.

[35] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[36] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug. 2009.

[37] D. W. Loehlin and S. B. Carroll. Expression of tandem gene duplicates is often greater than twofold. *Proc. Natl. Acad. Sci. U. S. A.*, 113(21):5988–5992, May 2016.

[38] M. Long, N. W. VanKuren, S. Chen, and M. D. Vibranovski. New gene evolution: little did we know. *Annu. Rev. Genet.*, 47:307–333, Sept. 2013.

[39] K. L. Mack, P. Campbell, and M. W. Nachman. Gene regulation and speciation in house mice. *Genome Res.*, 26(4):451–461, Apr. 2016.

[40] T. Maier, M. Güell, and L. Serrano. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.*, 583(24):3966–3973, Dec. 2009.

[41] T. N. Marriage, E. G. King, A. D. Long, and S. J. Macdonald. Fine-mapping nicotine resistance loci in drosophila using a multiparent advanced generation inter-cross population. *Genetics*, 198(1):45–57, Sept. 2014.

[42] B. Mcclintock. Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.*, 21:197–216, 1956.

[43] C. J. McManus, J. D. Coolon, M. O. Duff, J. Eipper-Mains, B. R. Graveley, and P. J. Wittkopp. Regulatory divergence in drosophila revealed by mRNA-seq. *Genome Res.*, 20(6):816–825, June 2010.

[44] C. J. McManus, G. E. May, P. Spealman, and A. Shteyman. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.*, 24(3):422–430, Mar. 2014.

[45] B. P. H. Metzger, P. J. Wittkopp, and J. D. Coolon. Evolutionary dynamics of regulatory changes underlying gene expression divergence among saccharomyces species. *Genome Biol. Evol.*, 9(4):843–854, Apr. 2017.

[46] S. Ohno. *Evolution by Gene Duplication*. Springer, Berlin, Heidelberg, 1970.

[47] S. Ohno. So much "junk" DNA in our genome. *Brookhaven Symp. Biol.*, 23:366–370, 1972.

[48] A. A. Pai and Y. Gilad. Comparative studies of gene regulatory mechanisms. *Curr. Opin. Genet. Dev.*, 29:68–74, Dec. 2014.

[49] S. Picelli, O. R. Faridani, A. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg. Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, 9(1):171–181, Jan. 2014.

[50] R Foundation for Statistical Computing, Vienna, Austria. R core team (2018). r: A language and environment for statistical computing. `https://www.r-project.org/`. Accessed: NA-NA-NA.

[51] G. Renaud, U. Stenzel, T. Maricic, V. Wiebe, and J. Kelso. deML: robust demultiplexing of illumina sequences using a likelihood-based approach. *Bioinformatics*, 31(5):770–772, Mar. 2015.

[52] B. Rhoné, C. Mariac, M. Couderc, C. Berthouly-Salazar, I. S. Ousseini, and Y. Vigouroux. No excess of Cis-Regulatory variation associated with intraspecific selection in wild pearl millet (cenchrus americanus). *Genome Biol. Evol.*, 9(2):388–397, Feb. 2017.

[53] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, Nov. 2007.

[54] M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, Apr. 2008.

[55] G. Rodrigo and M. A. Fares. Intrinsic adaptive value and early fate of gene duplication revealed by a bottom-up approach. *Elife*, 7, Jan. 2018.

[56] R. L. Rogers, L. Shao, and K. R. Thornton. Tandem duplications lead to novel expression patterns through exon shuffling in drosophila yakuba. *PLoS Genet.*, 13(5):e1006795, May 2017.

[57] I. G. Romero, I. Ruvinsky, and Y. Gilad. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.*, 13(7):505–516, June 2012.

[58] B. Schaefke, J. J. Emerson, T.-Y. Wang, M.-Y. J. Lu, L.-C. Hsieh, and W.-H. Li. Inheritance of gene expression level and selective constraints on trans- and cis-regulatory changes in yeast. *Mol. Biol. Evol.*, 30(9):2121–2133, Sept. 2013.

[59] N. J. Schurch, P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–851, June 2016.

[60] M. D. Shapiro, M. E. Marks, C. L. Peichel, B. K. Blackman, K. S. Nereng, B. Jónsson, D. Schluter, and D. M. Kingsley. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, 428(6984):717–723, Apr. 2004.

[61] X. Shi, D. W.-K. Ng, C. Zhang, L. Comai, W. Ye, and Z. J. Chen. Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in arabidopsis allopolyploids. *Nat. Commun.*, 3:950, July 2012.

[62] S. A. Signor and S. V. Nuzhdin. The evolution of gene expression in cis and trans. *Trends Genet.*, 34(7):532–544, July 2018.

[63] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, Oct. 2015.

[64] A. B. Singleton, M. Farrer, J. Johnson, A. Singleton, S. Hague, J. Kachergus, M. Hulihan, T. Peuralinna, A. Dutra, R. Nussbaum, S. Lincoln, A. Crawley, M. Hanson, D. Maraganore, C. Adler, M. R. Cookson, M. Muenter, M. Baptista, D. Miller, J. Blancato, J. Hardy, and K. Gwinn-Hardy. alpha-synuclein locus triplication causes parkinson's disease. *Science*, 302(5646):841, Oct. 2003.

[65] P. D. Sniegowski, P. G. Dombrowski, and E. Fingerman. Saccharomyces cerevisiae and saccharomyces paradoxus coexist in a natural woodland site in north america and display different levels of reproductive isolation from european conspecifics. *FEMS Yeast Res.*, 1(4):299–306, Jan. 2002.

[66] N. M. Springer and R. M. Stupar. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *Plant Cell*, 19(8):2391–2402, Aug. 2007.

[67] K. R. Takahasi, T. Matsuo, and T. Takano-Shimizu-Kouno. Two types of cis-trans compensation in the evolution of transcriptional regulation. *Proc. Natl. Acad. Sci. U. S. A.*, 108(37):15276–15281, Sept. 2011.

[68] I. Tirosh, S. Reikhav, A. A. Levy, and N. Barkai. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, 324(5927):659–662, May 2009.

[69] I. Tirosh, A. Weinberger, M. Carmi, and N. Barkai. A genetic signature of interspecies variations in gene expression. *Nat. Genet.*, 38(7):830–834, July 2006.

[70] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11):e112963, Nov. 2014.

[71] R. M. Waterhouse, M. Seppey, F. A. Simão, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, and E. M. Zdobnov. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*, Dec. 2017.

[72] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer-Verlag New York, 1 edition, 2009.

[73] P. J. Wittkopp, B. K. Haerum, and A. G. Clark. Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995):85–88, July 2004.

[74] P. J. Wittkopp, B. K. Haerum, and A. G. Clark. Regulatory changes underlying expression differences within and between drosophila species. *Nat. Genet.*, 40(3):346–350, Mar. 2008.

[75] P. J. Wittkopp, E. E. Stewart, L. L. Arnold, A. H. Neidert, B. K. Haerum, E. M. Thompson, S. Akhras, G. Smith-Winberry, and L. Shefner. Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in drosophila. *Science*, 326(5952):540–544, Oct. 2009.

[76] G. A. Wray. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*, 8(3):206–216, Mar. 2007.

[77] J.-X. Yue, J. Li, L. Aigrain, J. Hallin, K. Persson, K. Oliver, A. Bergström, P. Coupland, J. Warringer, M. C. Lagomarsino, G. Fischer, R. Durbin, and G. Liti. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.*, 49(6):913–924, June 2017.

[78] X. Zhang and J. J. Emerson. Inferring compensatory evolution of cis- and trans-regulatory variation. *Trends Genet.*, 35(1):1–3, Jan. 2019.

[79] X. Zhang and J. J. Emerson. Inferring the genetic architecture of expression variation from replicated high throughput allele-specific expression experiments. July 2019.

[80] H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, and L. Wang. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, Apr. 2014.