Title
Statistical and Computational Methods for Analyzing High-Throughput Genomic Data

Author
Li, Jingyi

Publication Date
2013

Peer reviewed|Thesis/dissertation

Statistical and Computational Methods for Analyzing High-Throughout Genomic Data

by

Jingyi Li

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter J. Bickel, Chair
Professor Haiyan Huang
Professor Sandrine Dudoit
Professor Steven E. Brenner

Spring 2013

Statistical and Computational Methods for Analyzing High-Throughput Genomic Data

Abstract

Statistical and Computational Methods for Analyzing High-Throughput Genomic Data

by

Jingyi Li

Doctor of Philosophy in Biostatistics
and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Peter J. Bickel, Chair

In the burgeoning field of genomics, high-throughput technologies (e.g. microarrays, next-generation sequencing and label-free mass spectrometry) have enabled biologists to perform global analysis on thousands of genes, mRNAs and proteins simultaneously. Extracting useful information from enormous amounts of high-throughput genomic data is an increasingly pressing challenge to statistical and computational science. In this thesis, I will address three problems in which statistical and computational methods were used to analyze high-throughput genomic data to answer important biological questions.

The first part of this thesis focuses on addressing an important question in genomics: how to identify and quantify mRNA products of gene transcription (i.e., isoforms) from next-generation mRNA sequencing (RNA-Seq) data? We developed a statistical method called Sparse Linear modeling of RNA-Seq data for Isoform Discovery and abundance Estimation (SLIDE) that employs probabilistic modeling and L_1 sparse estimation to answer this question. SLIDE takes exon boundaries and RNA-Seq data as input to discern the set of mRNA isoforms that are most likely to present in an RNA-Seq sample. It is based on a linear model with a design matrix that models the sampling probability of RNA-Seq reads from different mRNA isoforms. To tackle the model unidentifiability issue, SLIDE uses a modified Lasso procedure for parameter estimation. Compared with existing deterministic isoform assembly algorithms, SLIDE considers the stochastic aspects of RNA-Seq reads in exons from different isoforms and thus has increased power in detecting more novel isoforms. Another advantage of SLIDE is its flexibility of incorporating other transcriptomic data into its model to further increase isoform discovery accuracy. SLIDE can also work downstream of other RNA-Seq assembly algorithms to integrate newly discovered genes and exons. Besides isoform discovery, SLIDE sequentially uses the same linear model to estimate the abundance of discovered isoforms. Simulation and real data studies show that SLIDE performs as well as or better than major competitors in both isoform discovery and abundance estimation.

The second part of this thesis demonstrates the power of simple statistical analysis in correcting biases of system-wide protein abundance estimates and in understanding the rela-

tionship between gene transcription and protein abundances. We found that proteome-wide surveys have significantly underestimated protein abundances, which differ greatly from previously published individual measurements. We corrected proteome-wide protein abundance estimates by using individual measurements of 61 housekeeping proteins, and then found that our corrected protein abundance estimates show a higher correlation and a stronger linear relationship with mRNA abundances than do the uncorrected protein data. To estimate the degree to which mRNA expression levels determine protein levels, it is critical to measure the error in protein and mRNA abundance data and to consider all genes, not only those whose protein expression is readily detected. This is a fact that previous proteome-widely surveys ignored. We took two independent approaches to re-estimate the percentage that mRNA levels explain in the variance of protein abundances. While the percentages estimated from the two approaches vary on different sets of genes, all suggest that previous protein-wide surveys have significantly underestimated the importance of transcription.

In the third and final part, I will introduce a modENCODE (the Model Organism ENCyclopedia Of DNA Elements) project in which we compared developmental stages, tissues and cells (or cell lines) of *Drosophila melanogaster* and *Caenorhabditis elegans*, two well-studied model organisms in developmental biology. To understand the similarity of gene expression patterns throughout their developmental time courses is an interesting and important question in comparative genomics and evolutionary biology. The availability of modENCODE RNA-Seq data for different developmental stages, tissues and cells of the two organisms enables a transcriptome-wide comparison study to address this question. We undertook a comparison of their developmental time courses and tissues/cells, seeking commonalities in orthologous gene expression. Our approach centers on using stage/tissue/cell-associated orthologous genes to link the two organisms. For every stage/tissue/cell in each organism, its associated genes are selected as the genes capturing specific transcriptional activities: genes highly expressed in that stage/tissue/cell but lowly expressed in a few other stages/tissues/cells. We aligned a pair of *D. melanogaster* and *C. elegans* stages/tissues/cells by a hypergeometric test, where the test statistic is the number of orthologous gene pairs associated with both stages/tissues/cells. The test is against the null hypothesis that the two stages/tissues/cells have independent sets of associated genes. We first carried out the alignment approach on pairs of stages/tissues/cells within *D. melanogaster* and *C. elegans* respectively, and the alignment results are consistent with previous findings, supporting the validity of this approach. When comparing fly with worm, we unexpectedly observed two parallel collinear alignment patterns between their developmental timecourses and several interesting alignments between their tissues and cells. Our results are the first findings regarding a comprehensive comparison between *D. melanogaster* and *C. elegans* time courses, tissues and cells.

Dedicated to my family, in particular my mother Bo Yu and my grandmother Yanning Zhang

Contents

List of Figures

List of Tables

Acknowledgments

During my long journey of Ph.D. study at Berkeley, I received countless help from my advisors, collaborators, family and friends.

First and foremost, I am extremely grateful to my advisors Prof. Peter J. Bickel and Prof. Haiyan Huang for the scientific guidance and career mentorship they have generously provided. Since I joined their group, they have led me into the interdisciplinary field of statistical genomics and introduced me to the cutting edge research in the field. The maximum freedom and support they offered has allowed me to explore my scientific ideas, choose my research directions and develop my academic career. Working with Peter and Haiyan has been a joyful process filled with excitement and sense of accomplishment.

I also want to thank the other two professors on my dissertation committee, Prof. Sandrine Dudoit and Prof. Steven E. Brenner, for their insightful comments on my thesis. Steven also advised me on the *D. melanogaster*-*C. elegans* timecourse/tissue/cell comparison project (Chapter 4), and I greatly appreciate his guidance and scientific advice. Several people in the modENCODE consortium, including Dr. Susan E. Celniker, Dr. Roger Hoskins (both in LBNL), Prof. Robert Waterston (U Washington), Dr. LaDeana Hillier (WUSTL) Prof. Mark B. Gerstein (Yale), provided me with necessary information and advice needed in this project, and I strongly appreciate their help.

I would also like to thank my collaborator Dr. Mark D. Biggin in Lawarence Berkeley National Laboratories. We collaborated on three research projects including the correction of system-wide protein abundance estimates and re-determination of transcriptional importance (Chapter 3). Mark's biological expertise, scientific persistence and intuition taught me a good lesson of how to identify and address interesting biological questions.

I appreciate the numerous help I received from the faculty members in the Department of Statistics at UC Berkeley, especially Prof. Bin Yu, Prof. Cari Caufman and Prof. David Aldous. I am also grateful to members and former members in Bickel and Huang's group, in particular Dr. James B. Brown, Nathan Boley, and our former postdocs Dr. Ci-Ren Jiang, Dr. Qunhua Li and Dr. Hao Xiong, for their frequent discussions and comments on my research projects during our group meetings and daily conversations.

Finally, I would like to especially thank my family members and friends for their love and support. My father Youqiang Li, mother Bo Yu, grandma Yanning Zhang, and aunt Jin Yu have constantly given me strength and courage to pursue my dream of becoming a scientist all the way from my childhood. My boyfriend Henry Yitong Wu and my best friends (Jing Lu, Yao Zeng, Bingting Wen, Xiaoran Li, Han Chen, Zhenke Liu, Kui Qian, just to name a few) are also an indispensable part of my life and vital to my career achievements.

Jingyi (Jessica) Li

Contact Information

367 Evans Hall
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720 USA

Phone: (510) 847-1150
E-mail: jli@stat.berkeley.edu
WWW: www.stat.berkeley.edu/~jli

Research Interests

Developing statistical methods for biological datasets (e.g., next-generation RNA sequencing data); using statistics to understand scientific problems (e.g., mRNA isoform discovery and abundance estimation, cis-regulatory module identification, relationship between transcription and translation, comparison of developmental stages of multi-species); non-parametric regression; high-dimensional statistics

Education

- **University of California, Berkeley**, Berkeley, California USA

  Ph.D. Candidate, Biostatistics, since August 2008
  Designated Emphasis in Computational Biology
  - Dissertation Topic: "Statistical and Computational Methods for Analyzing High-throughput Biological Data"
  - Advisors: Peter J. Bickel and Haiyan Huang

- **Tsinghua University**, Beijing, China

  B.S., Biological Sciences (**summa cum laude**), June, 2007
  Minor in English, June, 2007
  - Overall GPA: 3.96/4.00; major GPA: 4.00/4.00
  - Thesis Topic: "*Oct4* is a target gene of *Wnt* signaling pathway in mouse ES cells"
  - Advisor: Duanqing Pei

Honors and Awards

- **Chinese Government Award for Outstanding Self-financed Students Aboard**, China Scholarship Council, 2013.
- **ISCB Travel Fellowship for RECOMB 2013** (17th Annual International Conference on Research in Computational Molecular Biology) at Tsinghua University, Beijing, China, 2013.
- **Stipend Award in Recognition of Scholastic Achievements**, Division of Biostatistics, UC Berkeley, 2013
- **International Dissertation Field Work Grant**, UC Berkeley Institute of International Studies, 2012-2013
- **Stipend Award in Recognition of Scholastic Achievements**, Division of Biostatistics, UC Berkeley, 2012
- **Best Presentation Award – CSHA Fellowship**, Cold Spring Harbor Asia Conferences: Bioinformatics of Human and Animal Genomics, Suzhou, China, Nov 2011
- **Funding for Participation in the Long-term Program "Mathematical and Computational Approaches in High-Throughput Genomics"**, IPAM (Institute of Pure and Applied Mathematics), Sep-Dec, 2011
- **Stipend Award in Recognition of Scholastic Achievements**, Division of Biostatistics, UC Berkeley, 2011
- **Outstanding Graduate Student Instructor Award**, UC Berkeley, 2010-2011
- **Distinguished College Graduate of Beijing**, 2007
- **Outstanding Undergraduate Thesis**, Tsinghua University, 2007
- **Distinguished Graduate of Class 2007**, Tsinghua University, 2007
- **Role-Model College Student of Beijing**, 2006
- **Merit-based "12.9" Fellowship** (awarded to $\sim$ top 30 students on campus regardless of individual majors), Tsinghua University, 2006
- **"Global Leadership" Finalist**, Unesco (United Nations Educational, Scientific and Cultural Organization), nominated by Tsinghua University, 2005
- **Merit-based Fellowship** (awarded to top 1% student in each department), Tsinghua University, 2005
- **Merit-based Fellowship** (awarded to top 1% student in each department), Tsinghua University, 2004

Teaching

- **University of California, Berkeley**, Berkeley, California USA **August, 2008 - present**

  Graduate Student Instructor

  Duties included shared administrative responsibilities with faculty instructors, office hours, weekly discussion sections, and grading.
  - STAT 215A (PhD course "Statistical Models: Theory and Application"; Instructor: Prof. Bin Yu), Department of Statistics, Fall 2012
  - STAT 210A (PhD course "Theoretical Statistics"; Instructor: Prof. Haiyan Huang), Department of Statistics, Fall 2010
  - STAT 200B (MA course "Introduction to Probability and Statistics at an Advanced Level"; Instructor: Prof. Cari Caufman), Department of Statistics, Spring 2010
  - STAT 131A (Upper level undergraduate course "Statistical Inferences for Social and Life Scientists"; Instructor: Prof. Haiyan Huang), Department of Statistics, Spring 2009

Academic Experience

- **IPAM (Institute of Pure and Applied Mathematics)** **Sep-Dec 2011**

  Participant in the long-term program "Mathematical and Computational Approaches in High-Throughput Genomics"

- **ICSA (International Chinese Statistical Association)** **2009**

  Member and volunteer in ICSA 2009 Applied Statistics Symposium

Publications

- **Li J.J.**, Bickel P.J., and Biggin M.D., "System wide analyses have underestimated protein abundances and transcriptional importance in animals", submitted to *Integrative Biology* and *ArXiv* `http://arxiv.org/abs/1212.0587`, 2012.

- **Li J.J.**, Huang H., Qian M, and Zhang X, "Transcriptome analysis using next-Generation sequencing", a chapter to appear in "*Advanced Medical Statistics*" (2nd Ed.), 2012.

- Fisher, W.W., **Li, J.J.**, Hammonds, A.S., Brown, J.B., Pfeiffer, B.D., Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J.A., Eisen, M.B. etc., "DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila", *Proceedings of the National Academy of Sciences* 109(52):21330–21335, 2012. `http://www.pnas.org/content/early/2012/12/05/1209589110.short`.

- The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome", *Nature* 489(7414):57–74, 2012. `http://www.nature.com/nature/journal/v489/n7414/full/nature11247.html`.

- Gao, Q. and Ho, C. and Jia, Y. and **Li, J.J.** and Huang, H., "Biclustering of linear patterns in gene expression data", 2012. *Journal of Computational Biology* 19(6):619–631, 2012. `http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0032`.

- **Li, J.J.** and Jiang, C.R. and Brown, J.B. and Huang, H. and Bickel, P.J., "Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation", *Proceedings of the National Academy of Sciences* 108(50):19867–19872, 2011. `http://www.pnas.org/content/108/50/19867.full`.

- MacArthur, S.*****, Li, X.Y.*****, **Li, J.***, Brown, J.B., Chu, H.C., Zeng, L., Grondona, B.P., Hechmer, A., Simirenko, L., Keränen, S.V., and etc., *Genome Biol* 10(7):R80, 2009. `http://www.biomedcentral.com/content/pdf/gb-2009-10-7-r80.pdf` (highly accessed article on BioMed Central; selected for Faculty of 1000 Biology and evaluated by Dr. Gregory Wray: see `http://www.f1000biology.com/article/id/1164185`) *** joint first authors**.

Papers in preparation

1. **Li J.J.**, Huang H., Bickel P.J., and Brenner S.E., "Comparison of *D. melanogaster* and *C. elegans* developmental stages by modENCODE RNA-Seq data."

2. **Li J.J.**, Bickel P.J., Zhang S., and Huang H., "Joint modeling of multiple samples for mRNA isoform discovery and abundance estimation from RNA-Seq data."

Conference Presentations

1. UC Systemwide Bioengineering Symposium, Berkeley, June 2012. Talk title: "Sparse linear modeling of RNA-seq data for isoform discovery and abundance estimation."

2. Joint mod/mouse/ENCODE AWG/PI meeting, MIT, May 2012. Talk title: "Developmental stage comparison of D. melanogaster and C. elegans and tissue and cell line comparison of D. melanogaster and H. sapiens with modENCODE/ENCODE RNA-Seq data."

3. Bay Area RNA Club, UCSF, Jan 2012. Talk title: "Sparse linear modeling of RNA-seq data for isoform discovery and abundance estimation."

4. Cold Spring Harbor Asia Conferences: Bioinformatics of Human and Animal Genomics, Suzhou, China, Nov 2011. Talk title: "Sparse linear modeling of RNA-seq data for

isoform discovery and abundance estimation" (selected as the Best Presentation Award – CSHA Fellowship).

5. Joint mod/ENCODE Consortia Meeting, Washington DC, May 2011. Talk and poster title: "Comparison between Developmental Stages of D. melanogaster and C. elegans with RNA-Seq data."

Invited Talks

1. *Statistical Methods for Analyzing High-throughput Genomic Data*, Department of Statistics, University of California at Davis, CA, January 16, 2013.

2. *Statistical Methods for Analyzing High-throughput Genomic Data*, Departments of Human Genetics and Statistics, University of California at Los Angeles, CA, February 1, 2013.

3. *Statistical Methods for Analyzing High-throughput Genomic Data*, Department of Statistics, University of Chicago, IL, February 25, 2013.

4. *Statistical Methods for Analyzing High-throughput Genomic Data*, Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Science, April 11, 2013.

Chapter 1

Introduction

In the last two decades, high-throughput technologies such as SAGE [1] and DNA microarrays [2] have revolutionized molecular biology, genomics and medicine by enabling biologists to perform global analysis on the expression of tens of thousands of genes simultaneously. The more recently emerged next-generation sequencing (NGS) technologies, which have lower cost, higher accuracy and less restrictions, further opens up the possibility of a wide variety of large-scale genomic research and is transforming genomic science into personal genomics [3]. SAGE, microarray and NGS are different generations of high-throughput genomic data. In a broader sense, high-throughput genomic data can also refer to data produced by high-throughput transcriptomic technologies (e.g. Deep-RACE [4]) and large-scale proteomic technologies (e.g. label-free mass spectrometry [5]). How to extract useful information from enormous amounts of high-throughput genomic data in various types is an increasingly pressing challenge to statistical and computational science. Interdisciplinary fields such as biostatistics, statistical genomics, bioinformatics, and computational biology continue to evolve to face such a challenge and to answer biological questions arising from high-throughput genomic data.

1.1 Examples of High-Throughput Genomic Data

First, we introduce two important types of high-throughput genomic data related to the studies to be addressed in this thesis.

1.1.1 RNA-Seq

In the burgeoning field of genomics, one of the most attractive research topic is to utilize next generation RNA sequencing (RNA-Seq data) for transcriptomic analysis. The RNA-Seq technology aims to capture RNA content of a biological sample by indirectly sequencing cDNAs reversely transcribed from extracted RNAs [6]. Advantages of RNA-Seq compared to previous microarray methods include its deep coverage, base-level resolution and no reliance

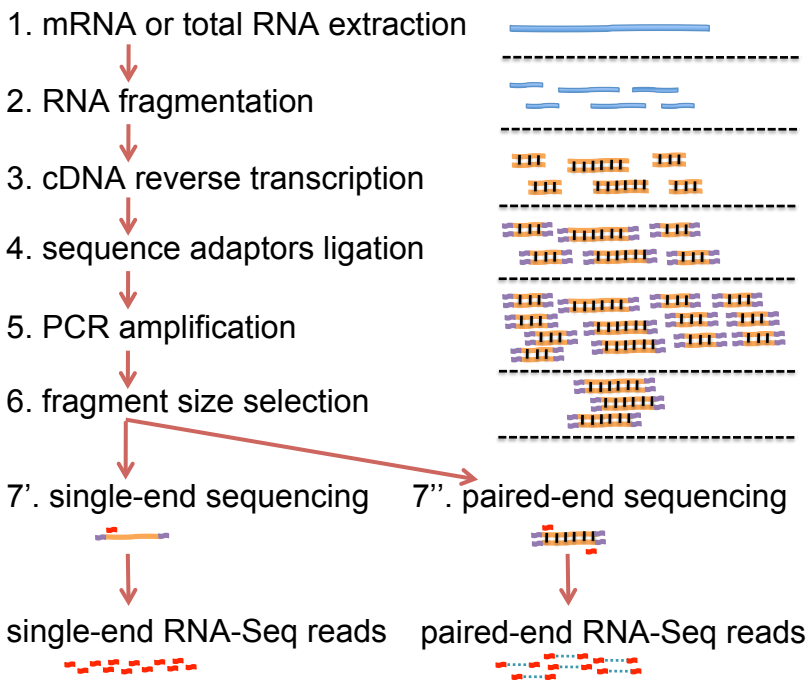

Figure 1.1: Data generation steps of a typical RNA-Seq assay.

on prior probe design or existing RNA transcript annotations. Hence, RNA-Seq is a key milestone in understanding the complex landscape and dynamics of eukaryotic transcriptomes. RNA-Seq can be used to simultaneously identify novel genes/transcripts, alternative splicing events, and rare genetic variants in a biological sample.

The sketch of a typical RNA-Seq assay is shown in Figure 1.1, which is derived from a review article on next-generation transcriptome assembly [7]. The first step is to extract RNAs (light blue) that are to be sequenced, followed by the second step of fragmenting extracted RNAs into short fragments. In the third step, the fragments will be reversely transcribed into cDNAs (yellow), which will then be ligated with sequencing adaptors in the fourth step. In the fifth step, PCR will be used to amplify the ligated cNDA fragments to ensure they have enough concentration. The sixth step is fragment size selection. In the final step, NGS is applied to sequence the ends of the retained cDNA fragments. The two ends of a single cDNA fragment will be treated as unrelated reads if single-end sequencing protocol is used; on the other hand, paired-end sequencing will keep the pairing relationship of the two ends. RNA-Seq read length varies from 25 base pairs (bp) to 400 bp, depending on different platforms and protocols.

RNA-Seq data (i.e., reads) represent quantities of short sequences in existing RNA transcripts, and therefore can be used to discover and quantify RNA transcripts in a biological sample.

1.1.2 Label-Free Mass Spectrometry

System-wide quantification of protein abundance in a biological sample has been a long-lasting problem in proteomics and biochemistry. Mass-spectrometry-based techniques have made phenomenal impact on protein identification and quantification. However, to date, identification and quantification of all the proteins in a biological sample remains an unmet technical challenge [5]. Nevertheless, label-free mass-spectrometry quantification strategies provide a practical way of whole proteome quantification, at a sacrifice of accuracy though. Label-free mass spectrometry has a simple workflow that skips traditional mass-spectrometry steps including protein/peptide labeling and purification/fractionation [5]. Despite its reduced accuracy in protein quantification compared to traditional labeling mass-spectrometry methods, label-free mass-spectrometry provides valuable resources for understanding system-wide translation and studying differential protein expression between biological samples.

1.2 Example Questions to be Addressed by Statistical Analysis on High-Throughput Genomic Data

Next, we discuss several important biological questions that require using statistical and computational methods to analyze high-throughput genomic data.

Questions in the genomics field can be summarized around the central dogma. In the DNA state, questions include transcription factor binding sites and intensities, DNA methylation, histone modification, etc. In the RNA state, questions can be asked about mRNA isoforms (i.e., different mRNA transcripts arising from the same gene), alternative splicing, gene expression, etc. In the protein state, the most straightforward question is system-wide protein quantification. More questions can be asked about the transition processes between the three states, i.e., transcription (connecting DNA and RNA states) and translation (connecting RNA and protein states).

For example, we have used statistical analysis to determine the quantitative relationship between transcription factor binding and downstream gene expression, and found that the relationship is both discrete and continuous: transcription factor binding below a certain threshold does not trigger specific gene expression, while the binding intensity beyond that threshold has a strong correlation with known biological and transcriptional regulatory specificities [8, 9].

In this thesis, we will address three important questions, where the first and the third questions are related to the RNA state and the second question is about transcription and translation.

1.2.1 Discovery and Quantification of RNA Isoforms from RNA-Seq Data

Unlike previous technologies such as microarrays, RNA-Seq provides novel splice junction information in addition to gene expression, thus facilitating assembling full-length mRNA isoforms (mRNA molecules transcribed from the same gene but having different sequences) (i.e., "isoform discovery") and quantifying isoform expressions (i.e., "isoform abundance estimation"). Before the invention of RNA-Seq, mRNA isoforms have been discovered on a gene to gene basis and recorded in annotations (databases of reported genes and their mRNA isoforms). Previous methods for isoform discovery and abundance estimation from RNA-Seq data can be divided into two categories: "annotation-based" and "annotation-free". The former takes existing annotations to define genes and isoforms; the latter uses no annotation information but directly assembles mRNA isoforms from RNA-Seq data. Methods in either category have their drawbacks, because existing annotations are incomplete and RNA-Seq data contain various noise and biases.

In Chapter 2, we developed a method entitled "SLIDE" (Sparse Linear modeling of RNA-Seq data for Isoform Discovery and abundance Estimation), which defines a new category—"annotation-aided" methods. SLIDE is an annotation-aided method that lies in the middle of the annotation usage spectrum—from completely ignoring to totally depending on annotations—and combines benefits from both ends [10]. SLIDE has the advantage of utilizing both literature and data information to find potential novel isoforms (compared to annotation-based methods) without being biased by RNA-Seq data noise in defining gene and exon boundaries (compared to annotation-free methods). If supplemented with de novo genome assemblies from other RNA-Seq software packages, SLIDE can also discover isoforms involving novel genes/exons.

1.2.2 System-wide Protein Quantification and the Importance of Transcription in Determining Protein Abundance

Label-free mass spectrometry methods have recently been developed to determine the absolute number of protein molecules per cell for thousands of genes. Because the methods are known to have lower accuracy than previous labeling mass spectrometry and other smaller-scale quantification methods, it is necessary to ask whether the system-wide protein abundance estimates by label-free quantification are accurate. Another question is to estimate the importance of transcription in determining protein levels given the system-wide estimates of mRNA and protein abundances.

In Chapter 3, we found that the published proteome-wide surveys have significantly underestimated protein abundances. Aiming to use statistics to correct the bias in those protein abundance estimates, we re-analyzed the system-wide protein mass spectrometry data in Schwanhausser et al [11] against previous individual protein abundance measurements, and corrected the system-wide data by fitting a two-part spline model. After our correction, we observed significantly improved correlation between protein abundance estimates and mRNA

levels. We also found that transcription contributes a higher percentage to the variance of protein levels than Schwanhausser et al estimated, and transcription contributes more than translation does to the variance of protein levels, in contrast to what Schwanhausser et al claimed. Our results [12] raised a caution about systems biology modeling without proper data scaling or thorough accounting for experimental errors.

1.2.3 Comparison of Biological Samples from Different Species by Gene Expression

Given system-wide gene expression estimates from high-throughput genomic data (e.g. RNA-Seq), can we compare two biological samples from different species in terms of their transcriptional similarity?

In Chapter 4, we addressed this question in the context of two model organisms, *D. melanogaster* and *C. elegans*. The production of modENCODE [13, 14, 15] RNA-Seq data at different developmental stages, tissues and cells (or cell lines) of *D. melanogaster* (fly) and *C. elegans* (worm) enables a transcriptome-wide comparison study to understand the evolutionary conservation of developmental biology of the two model organisms. Our approach centers on using orthologous genes to link the two organisms, and employing system-wide gene expression estimates to find sample-associated genes to represent characteristic transcriptional events in every sample. A hypothesis testing approach similar to hypergeometric testing is developed to compare different samples in terms of overlap (for samples from the same species) or orthology (for samples from different species) in their associated genes. Interesting comparison results were found between developmental stages, tissues and cells within and between the two species. Our results are the first findings regarding a comprehensive comparison of various developmental stages, tissues and cells of *D. melanogaster* and *C. elegans*. The results and the comparison approach will make a valuable contribution to developmental biology and comparative genomics.

Chapter 2

Sparse Linear Modeling of Next-Generation mRNA Sequencing (RNA-Seq) Data for Isoform Discovery and Abundance Estimation

2.1 Introduction

The recently developed next-generation mRNA sequencing (RNA-Seq) assay, with deep coverage and base level resolution, has provided a view of eukaryotic transcriptomes of unprecedented detail and clarity. Unlike microarrays, RNA-Seq data have novel splice junction information in addition to gene expression, thus facilitating whole-transcriptome assembly and mRNA isoform quantification. RNA-Seq data includes both single-end and paired-end reads, where a single-end read is a sequenced end of a cDNA fragment from an mRNA transcript, and a paired-end read is a mate pair corresponding to both ends of a cDNA fragment.

In the mRNA isoform discovery field, one of the most widely used software packages is Cufflinks [16]. It builds a set of genes and exons solely from RNA-Seq data first, and subsequently uses a deterministic approach to find a minimal set of isoforms that can explain all the cDNA fragments indicated by paired-end reads. Cufflinks mainly uses qualitative exon expression and junction information in its isoform discovery, lacking a quantitative consideration of RNA-Seq data. Although Cufflinks gives very useful results, we note that the isoforms it discovers based on de novo assembled genes and exons can be heavily biased by different types of RNA-Seq data noise [17, 18, 19, 20]. Two recently published modENCODE (Model Organism Encyclopedia of DNA Elements) [13] consortium papers [14, 15] also raise concerns about relying solely on RNA-Seq reads in isoform discovery and have suggested using manual annotations to scrutinize the results.

In the mRNA isoform quantification field, the question is to estimate the abundance of

isoforms in a given set. Available abundance estimation methods include direct computation [21, 22] and model-based approaches. Many model-based studies [16, 23, 24, 25, 26] have used maximum-likelihood approaches to estimate isoform abundance. There are also efforts on formulating the abundance estimation problem as a linear model [27], where the independent and dependent variables are isoform expression levels and categorized RNA-Seq read counts, respectively. In particular, binary values have been used in the design matrix to relate categorized reads to different isoforms, but that design matrix misses the quantitative relationship between read quantities and isoform abundance.

In this study, we propose a statistical method called "Sparse Linear modeling of RNA-Seq data for Isoform Discovery and abundance Estimation" (SLIDE) that uses RNA-Seq data to discover mRNA isoforms given an extant annotation of gene and exon boundaries, and to estimate the abundance of the discovered or other specified mRNA isoforms. The extant annotation can come from annotation databases [e.g., Ensembl [28] or UCSC Genome Browser [29]], can be supplemented by other transcriptomic data such as RACE or CAGE (18, 19), or can even be inferred from RNA-seq de novo assembly algorithms [16, 30]. SLIDE is based on a linear model with a nonbinary design matrix modeling the sampling probability of RNA-Seq reads from mRNA isoforms. When modeling the design matrix, we considered the effects of GC content, cDNA fragment lengths, and read starting positions. This linear model, coupled with the carefully defined design matrix, gives SLIDE a stochastic property of making use of exon expression quantitatively in isoform discovery. The SLIDE model can also be easily extended to incorporate other transcriptomic data [e.g., RACE [31], CAGE [32], and EST [33]] with RNA-Seq to achieve more comprehensive results. The SLIDE software package is available at `https://sites.google.com/site/jingyijli/SLIDE.zip`.

2.2 Results

2.2.1 Linear Modeling for RNA-Seq Data

SLIDE is designed as a tool for discovering mRNA isoforms and estimating isoform abundance from RNA-Seq reads, on top of known information about gene and exon boundaries. For isoform discovery, SLIDE considers all the possible isoforms by enumerating exons of every gene. For example, a gene of n nonoverlapping exons has $2^n - 1$ possible isoforms, each composed of a subset of the n exons. However, because of the possible occurrence of alternative splicing within exons, isoforms of the same gene may have partially overlapping but different exons. Hence, for ease of enumeration, we define a subexon as a transcribed region between adjacent splicing sites in any annotated mRNA isoforms (Figure 2.1A). With this definition, every gene has a set of nonoverlapping subexons, from which we can enumerate all the possible isoforms including annotated ones.

We formulate the task of discovering isoforms for a given gene as a sparse estimation problem where the sparseness applies to the isoforms expected from RNA-Seq data. Because exon expression levels and the existence of possible exon-exon junctions are the key for

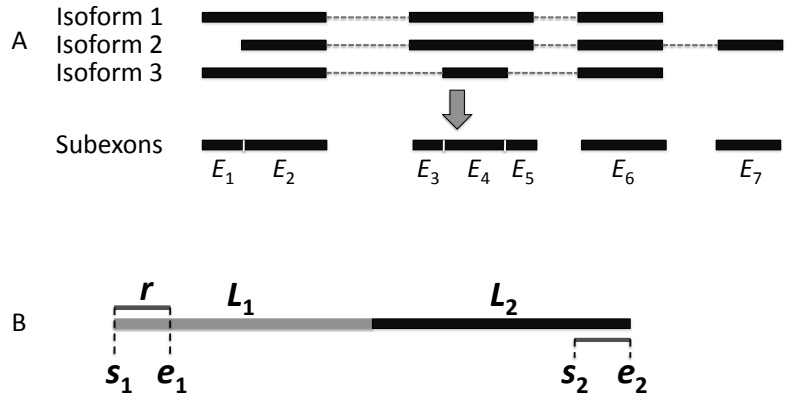

Figure 2.1: Definition of subexons and notations. (A) Subexons are defined as transcribed regions between adjacent alternative splicing sites. (B) A two-exon mRNA transcript. s_1, e_1, s_2, and e_2, genomic positions associated with a paired-end read. r, the read end length; L_1 and L_2, the exon lengths.

isoform discovery and they can be inferred from the starting and ending positions of RNA-Seq reads mapped to a reference genome, we are motivated to transform RNA-Seq reads into a summary that captures the key information. For a paired-end read, we exact four genomic locations s_1, e_1, s_2, and e_2, where s_1 and e_1 are the starting and ending positions of its 5' end, and s_2 and e_2 are the starting and ending positions of its 3' end (Figure 2.1B). Note that a paired-end read uniquely corresponds to a cDNA fragment with both ends sequenced, that is, s_1 and e_2 are the starting and ending positions of the fragment, respectively. We next categorize paired-end reads into *paired-end bins* defined as four-dimensional vectors: Bin (i, j, k, l) contains reads whose s_1, e_1, s_2 and e_2 are in subexons i, j, k and l respectively (see Subsection 2.4.1 for more detail). For single-end reads, we can similarly categorize them into two-dimensional *single-end bins*. The so-defined bin counts provide all the exon expression and junction information.

SLIDE is built upon a linear model whose design matrix $\mathbf{F}$ models conditional probabilities of observing reads in different bins given an isoform. For paired-end data, modeling $\mathbf{F}$ requires distributional assumptions on the two ends (i.e., s_1, e_2) of a cDNA fragment in an mRNA transcript, or equivalently on the fragments 5' end (i.e., s_1) and its length (i.e., $e_2 - s_1$). For s_1, uniform distribution assumptions have been widely used. However, after considering the high correlation observed between sequencing read coverage and genome GC content [17], we assume the density of s_1 is uniform within subexons and proportional to the GC content between subexons. We specify the distribution of the fragment length, $e_2 - s_1$, by assuming e_2 to follow a Poisson point process given s_1 fixed. Consequently, $e_2 - s_1$ is modeled as truncated Exponential after taking into account the size selection step in RNA-Seq protocols (see Subsection 2.4.2 for more detail). Another widely used fragment length distribution is Normal distribution [16], which is also implemented in SLIDE and compared

with truncated Exponential (see Appendix A.1.1).

We then use a linear model as approximation to the observed bin proportions,

$$b_j = \sum_{k=1}^{K} F_{jk} p_k + \epsilon_j, \quad j = 1, \cdots, J. \tag{2.1}$$

where b_j is the observed proportion of reads in the jth bin, $F_{jk} = \Pr(j\text{th bin} \mid k\text{th isoform})$ (i.e., the conditional probability of observing paired-end reads in the jth bin given that they are from the kth isoform), p_k is the proportion of the kth isoform to be estimated, and ϵ_j is the error term with mean 0. Besides, J and K are the numbers of bins and isoforms, respectively (see Subsection 2.4.1 for more detail). This is the core linear model used in SLIDE for both isoform discovery and abundance estimation of discovered isoforms. For isoform discovery, usually $K > J$, so the model is unidentifiable. But based on biological knowledge, we expect the model to be sparse and achieve sparse estimation by a modified Lasso [34] method (see Subsection 2.4.3 for more detail). For abundance estimation, only the proportions of discovered isoforms are parameters in the linear model, and their number is often far less than K, so there is no identifiability issue anymore. SLIDE then does the parameter estimation by nonnegative least squares. Compared with maximum-likelihood approaches used by other abundance estimation methods, SLIDE has the computational advantage of fitting a linear model as an intrinsic element.

2.2.2 Simulation Results

A simulation study is used to assess the accuracy of SLIDE on isoform discovery and abundance estimation. We simulated reads from genes and true mRNA isoforms extracted from *D. melanogaster* annotation (September 2010) of UCSC Genome Browser [29]. For illustration purposes, we focus on the 3,421 genes on chr3R. Based on our defined subexons, those genes consist of 34.2% with 1-2 subexons, 57.6% with 3-10 subexons, and 8.2% with more than 10 subexons. Because the estimation for genes with 1-2 subexons is trivial due to their small numbers of possible isoforms, and genes with more than 10 subexons only constitute a small proportion and their estimation is computationally costly, we applied SLIDE to the subset of 3-10 subexons, 1,972 genes in total. We generated 500×50 (runs) paired-end reads for each gene from annotated isoforms of randomly defined proportions, and then we applied SLIDE to the simulated reads for isoform discovery and abundance estimation.

The isoform discovery results of all 50 runs are in Figure 2.2A. We divided genes into groups by their numbers of subexons n ($n = 3, \cdots, 10$). For each gene, SLIDE returns a vector of estimated proportions of all its possible isofoms. We define isoforms whose estimated proportions exceed threshold 0.1 as *discovered isoforms* and evaluate them by the UCSC annotation. (Note that other thresholds 0.05 and 0.2 return similar results.) For each gene, the precision rate is defined as $TP/(TP + FP)$, and the recall rate is $TP/(TP + FN)$, where TP is the number of true positives (discovered isoforms that are also in the annotation), FP is the number of false positives (discovered isoforms that are not in the

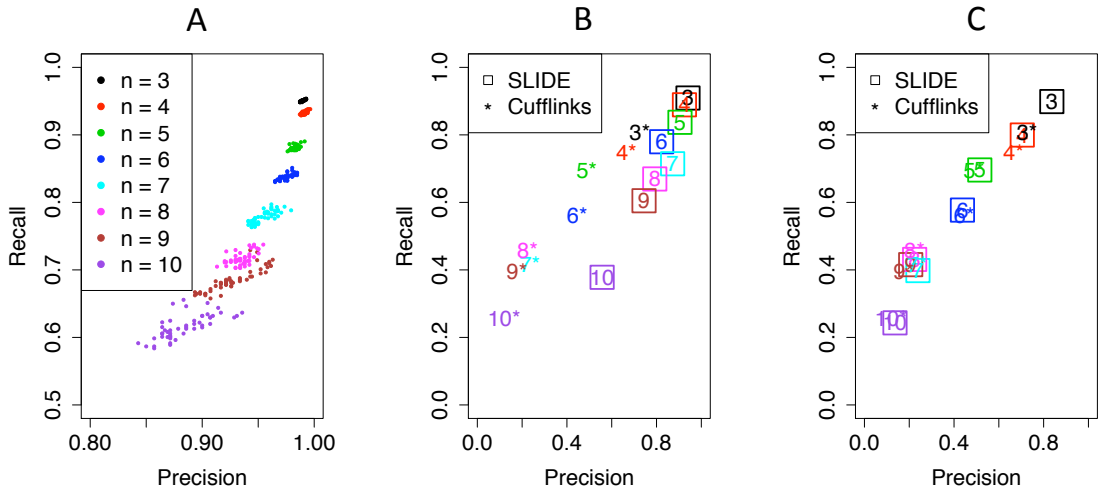

Figure 2.2: Isoform discovery results. (A) Precision and recall rates of SLIDE on 50 simulated datasets, with different colors for groups of genes with n subexons ($n = 3, \cdots, 10$) and every point representing the average precision and recall rates of every group on one dataset. (B) Precision and recall rates of SLIDE (using annotated genes/exons) and Cufflinks on dataset 1 (Table 2.1). Numbers, group indices of genes (i.e., numbers of subexons); squares/stars, SLIDE/Cufflinks results. (C) Precision and recall rates of SLIDE (using Cufflinks assembled genes/exons) and Cufflinks on dataset 1.

annotation), and FN is the number of false negatives (undiscovered isoforms that are in the annotation and have every exon observed). For each group of n-subexon genes, we calculated their average precision and recall rates as presented in Figure 2.2A. The results show that SLIDE maintains high precision rates ($> 80\%$) and good recall rates ($> 60\%$) in all groups of genes. In particular, for genes with three and four subexons, the precision and recall rates are greater than 98% and 92%, respectively. As n increases, the precision and recall rates decrease, and the variance between different simulation runs increases. This observation is reasonable because with the increase of n, the number of possible isoforms increases exponentially, as does the difficulty of isoform discovery.

To illustrate the abundance estimation accuracy of SLIDE, we applied it to 317 multi-isoform genes on chr3R in the UCSC annotation (798 isoforms in total), with the same simulated paired-end reads. From reads of each simulation run, SLIDE estimates the 798 isoform proportions normalized by each gene. We calculated the Pearson correlation between the estimates and the true isoform proportions used in the simulation, and we found that the correlation coefficients of the 50 runs range from 0.92 to 0.95. We also illustrate the abundance estimation accuracy of SLIDE by a scatter plot of the median estimated isoform proportions over the 50 runs vs. true isoform proportions in Figure 2.3A ($R = 0.99$).

This simulation study shows satisfactory performance of SLIDE in isoform discovery

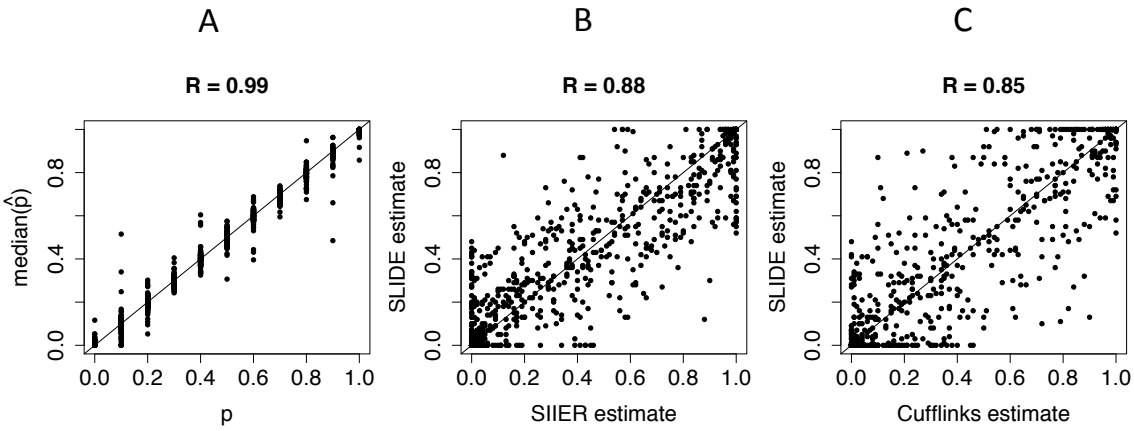

Figure 2.3: Abundance estimation results. (A) p vs. median($\hat{p}$) of 798 isoforms on 50 simulated datasets. p, true isoform proportion; median($\hat{p}$), median of the 50 estimated isoform proportions. (B) SLIDE vs. SIIER estimates of the 798 isoforms on dataset 1 (Table 2.1). (C) SLIDE vs. Cufflinks estimates of the 798 isoforms on dataset 1.

and abundance estimation. Further simulation studies with lowly expressed genes are in Appendix A.2.1.

2.2.3 mRNA Isoform Discovery on modENCODE Data

The main feature of SLIDE is discovery of mRNA isoforms from RNA-Seq data. Four modENCODE [13] *D. melanogaster* RNA-Seq datasets (Table 2.1) are used in the real data analysis. Again, for illustration purposes, we focus on the 1,972 genes with 3-10 subexons on chr3R of *D. melanogaster*. To avoid the effects of high false positive and negative rates of RNA-Seq data in lowly expressed genes [35], we applied SLIDE to genes with RPKM (number of reads per kilobase per million of mapped reads) [22] greater than 1.

We compare SLIDE with Cufflinks (version 0.9.3) in terms of their isoform discovery precision and recall rates, evaluated by the UCSC annotation in a similar way to the simulation study (see Appendix A.3.1). We note that SLIDE and Cufflinks target the isoform discovery problem from two different aspects. SLIDE discovers isoforms from given gene and exon structures, whereas Cufflinks contructs isoforms from its de novo assembled genes and exons. Hence, we carried out the comparison in two ways: (i) SLIDE with input genes and exons from the UCSC annotation vs. Cufflinks; (ii) SLIDE with input genes and exons assembled by Cufflinks vs. Cufflinks. The former is to evaluate the overall performance of the two methods under their default settings, whereas the latter is to specifically compare their isoform construction performance given the same set of genes and exons. The comparison

Dataset	Type	Sample	Read length	Total number of reads	Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) numbers
1	paired-end	ML-DmBG3-c2	37 bp	25,094,224	SRX003838, SRX003839
2	paired-end	Kc167	37 bp	18,602,220	SRX003836, SRX003837
3	paired-end	Kc167	76 bp	20,118,748	SRR070261, SRR070269, SRR111873
4	paired-end and single-end	embryo 16-17h	76 bp	23,388,810 and 27,913,445	SRR023600, SRR035402, SRR023720, SRR023715, SRR023751, SRR023707, SRR023826

Table 2.1: modENCODE datasets used in the analysis

results on dataset 1 (Table 2.1) are summarized in Figure 2.2B and C. (See Appendix A.3.1 for results on other datasets.)

Figure 2.2B, corresponding to the first comparison, shows that SLIDE with input genes and exons from the annotation has significantly higher precision and recall rates than Cufflinks. In the second comparison, with de novo genes and exons assembled by Cufflinks, SLIDE has better precision and recall rates than Cufflinks has for genes with three and four subexons, and for the rest of genes, the two methods have similar performance (Figure 2.2C). We observe that the overall precision and recall rates in Figure 2.2C are worse than those of SLIDE in Figure 2.2B. These results remind us of the concerns voiced by other researchers about constructing isoforms based on de novo genes and exons built solely from RNA-Seq data [14, 15]. We speculate that results of the second comparison are not enough to illustrate the isoform construction performance of SLIDE and Cufflinks, because the similarly low precision and recall rates observed in Figure 2.2C might have been dominated by the disagreement between the de novo assembled genes/exons and the annotation. Hence, we performed an additional comparison on a smaller set of 246 genes whose de novo exons assembled by Cufflinks agree with the annotation. This comparison provides a direct evaluation on the isoform construction performance of SLIDE and Cufflinks. We found that isoforms discovered by SLIDE have an average precision rate of 93% and a recall rate of 96%, both higher than the average precision rate (89%) and recall rate (94%) of isoforms found by Cufflinks. This result demonstrates that SLIDE has higher accurracy than Cufflinks has in isoform construction from a given set of genes and exons. For more details, see Appendix A.3.1.

By a detailed inspection of the isoforms discovered by Cufflinks, we find that many discovered isoforms are fragments of annotated isoforms in public databases. This is mainly due to the difficulty in de novo construction of gene boundaries. Cufflinks also has troubles in detecting lowly expressed genes de novo. By contrast, SLIDE can discover correct isoforms even with a small number of reads, based on existing gene boundary information. For instance, when applied to dataset 1, SLIDE has discovered isoforms in 1,084 genes (RPKM $>$ 1) out of the total 1,972 genes, whereas Cufflinks has only found isoforms in 801 genes. These observations confirm again the importance of having correct gene boundaries in isoform discovery. Another advantage of SLIDE is the usage of a stochastic approach to simultaneously detect isoforms with alternative starts/ends [e.g., (1,2,3,4) and (2,3,4)], where Cufflinks will only discover the longest one (1). However, when there are significant RNA-Seq data bi-

ases in 5' and 3' ends of mRNA transcripts, the deterministic approach of Cufflinks may be more robust. In the future, with the continuing development of sequencing technology and promising improvement in RNA-Seq signal-to-noise ratios, we would expect the stochastic approach of SLIDE to be preferred.

There are other isoform discovery methods that use sparse estimation but with different methodology [27, 36]. A numerical comparison between SLIDE and IsoLasso [27] shows that SLIDE has higher accuracy in isoform discovery. For detailed comparison information, please see Appendix A.3.2.

2.2.4 mRNA Isoform Abundance Estimation on modENCODE Data

Another feature of SLIDE is to estimate the abundance of mRNA isoforms discovered or other specified (e.g., annotated) from an RNA-Seq sample. Because of the lack of ground truth of isoform abundance in datasets 1-4 (Table 2.1), to evaluate the abundance estimation performance of SLIDE, we compare its estimates to those of two popular methods: statistical inferences for isoform expression in RNA-Seq (SIIER) [24] and Cufflinks [16]. Note that SLIDE returns estimates of mRNA isoform proportions that are equivalent and convertible to the common abundance measure, isoform RPKMs [22] used in SIIER.

In the comparison between SLIDE and SIIER, both methods estimate the isoform abundance of the 317 chr3R genes with multiple isoforms in the UCSC annotation. In dataset 1, after removing 25 genes with high expression variance among exons (see Appendix A.4), we obtain a scatter plot of the two sets of estimates in Figure 2.3B ($R = 0.88$). A similar comparison is carried out between SLIDE and Cufflinks, and the results are in Figure 2.3C ($R = 0.85$). The results show that SLIDE obtains estimates similar to those of SIIER and Cufflinks. For more discussions on the results, see Appendix A.4.

2.2.5 Miscellaneous Effects on Isoform Discovery

Using datasets 1-4 (Table 2.1), we study the following critical issues affecting isoform discovery from RNA-Seq data.

1. *GC content variation.* To study the usefulness of considering GC content variation in isoform discovery, we additionally implemented another version of $\mathbf{F}$, assuming the cDNA fragment starting position s_1 as uniform across all subexons. Note that our default $\mathbf{F}$ assumes the density of s_1 as uniform within subexons but proportional to GC content between subexons, as motivated by observed high correlation between read coverage and GC content variation (2, 4) (see Appendix A.6). Isoform discovery results on dataset 1 by SLIDE based on the two version of $\mathbf{F}$ are compared in Table 2.2. Recall rates are similar in both results, but precision rates are improved with the consideration of GC content. These results indicate that GC content can provide SLIDE with useful

n		3	4	5	6	7	8	9	10
without GC	precision	0.93	0.90	0.87	0.80	0.83	0.75	0.71	0.49
	recall	0.91	0.89	0.83	0.77	0.71	0.68	0.61	0.36
with GC	precision	0.94	0.92	0.90	0.82	0.87	0.79	0.74	0.56
	recall	0.91	0.89	0.84	0.78	0.71	0.67	0.60	0.38

Table 2.2: Comparison of isoform discovery results by SLIDE with two versions of **F**

information in modeling **F**, and thus support various attempts of using GC content information to correct RNA-Seq data noise [18, 19].

2. *Read/fragment length effects.* To explore the effects of RNA-Seq read lengths on isoform discovery, we applied SLIDE to datasets 2 and 3. The two datasets are generated from the same Kc167 sample of similar sequencing depth but with different read lengths: 37 bp (dataset 2) vs. 76 bp (dataset 3). We compare the isoform discovery results on both datasets in Figure 2.4A. The precision and recall rates for genes with 3-9 subexons are surprisingly higher with the 37-bp data than the 76-bp data. This result contradicts our expectation that RNA-Seq data with longer read length would provide more information on exon junctions that are crucial to isoform discovery. Trying to find a plausible explanation, we checked the empirical distribution of cDNA fragment lengths in single-exon genes for both data, and found the distribution close to $N(166, 26^2)$ and $N(127, 13^2)$ for the 37-bp and 76-bp data, respectively. The fact that the 37-bp data contain a greater number of long fragments is a result of different experimental protocols, and is likely to be a reason for the observed unexpected comparison results. A simulation study with different read and fragment lengths reveals that the fragment length distribution has larger effects than the read length has on isoform discovery, and to some extent confirms our real data observation (see Appendix A.5).

3. *Paired-end vs. single-end RNA-Seq data.* Compared with single-end RNA-Seq data, the more recent paired-end data provides more information on exon junctions and thus is expected to return isoform discovery results with higher precision rates. But if both single-end and paired-end data are available for the same RNA-Seq sample, the former can possibly complement the latter by providing more exon expression information, helping capture lowly expressed exons in rare isoforms, and thus resulting in isoform discovery results with higher recall rates. Because SLIDE has the flexibility of inputting both single-end and paired-end RNA-Seq data (see Subsection 2.4.2 for more detail), we tested these hypotheses by applying it to dataset 4, which has both single-end and paired-end data from the same sample and of similar numbers of reads (Table 2.1). We specifically compare the results of SLIDE on (i) paired-end data, (ii) single-end data, and (iii) both paired-end and single-end data in Figure 2.4B. From the figure, we observe that using paired-end data alone has the highest precision rates for all the

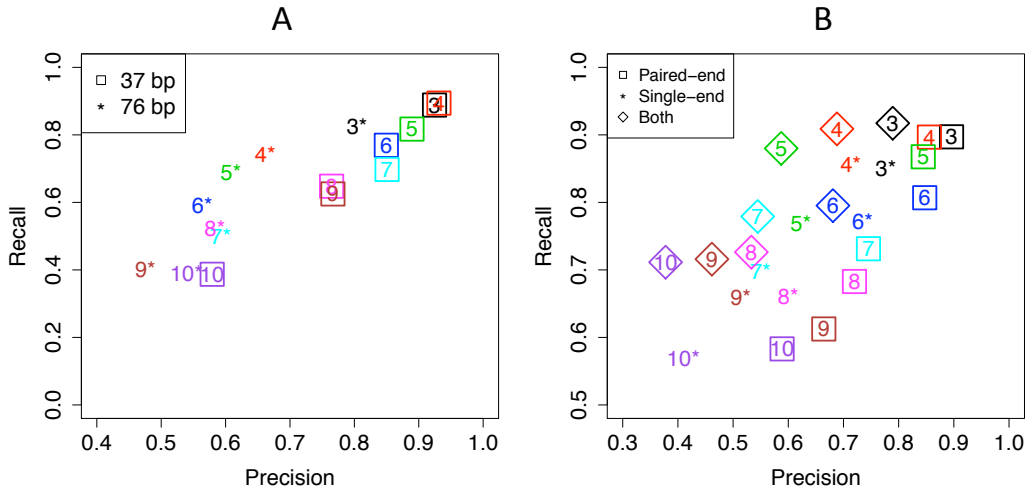

Figure 2.4: Miscellaneous effects. (A) Precision and recall rates of SLIDE on 37 bp and 76 bp paired-end RNA-Seq data (datasets 2-3). (B) Precision and recall rates of SLIDE on dataset 4 with paired-end data only (squares), single-end data only (stars), and both (diamonds).

genes, whereas using both data has the best recall rates. These results confirm our intuitive hypotheses that paired-end data alone gives more precise information than single-end data does in isoform discovery; however, single-end data does provide extra exon expression information as well as noise when it is used in addition to paired-end data, hence resulting in higher recall rates and lower precision rates.

2.3 Discussion

We have proposed a sparse linear model approach (SLIDE) capable of discovering mRNA isoforms of given genes and estimating the abundance of discovered or other specified isoforms from RNA-Seq data. Compared to existing approaches [16, 24], SLIDE (i) discovers isoforms from all possible ones based on known gene and exon boundaries (e.g., from the UCSC annotation), (ii) uses a stochastic approach with a quantitatively modeled design matrix $\mathbf{F}$ (i.e., conditional probabilities of observing RNA-Seq reads from mRNA isoforms) in isoform discovery, (iii) uses the same linear model subsequently for abundance estimation on discovered or other specified isoforms, and (iv) can be used as a downstream isoform discovery tool of de novo gene and exon assembly algorithms. Other widely used isoform discovery methods [16, 30] find isoforms based on their own de novo genes and exons solely assembled from RNA-Seq reads, and thus their discovered isoforms are highly dependent on the accuracy of de novo assembly. SLIDE can avoid possible de novo assembly errors [17] by using known gene and exon boundaries; it can also integrate de novo assemblies with known ones to prevent the risk of missing isoforms involving novel exons. SLIDE will also benefit

from ongoing efforts of improving *D. melanogaster* transcriptome annotations [13].

We have also explored various factors that may affect the performance of SLIDE on isoform discovery. Our results suggest that (i) the consideration of GC content variation in modeling $\mathbf{F}$ can improve the precision, (ii) the cDNA fragment size selection protocol and the resulting cDNA fragment lengths have larger effects than read lengths have on both the precision and recall, and (iii) paired-end RNA-Seq data provides more accurate information than single-end data does in isoform discovery, but the addition of single-end data would help with the discovery of rare isoforms.

As demonstrated by the isoform discovery and abundance estimation results, SLIDE shows great promise as a tool for handling the two tasks sequentially with a shared linear model. The modeled design matrix $\mathbf{F}$ is also shown to be a good quantitative representation of sampling RNA-Seq reads from mRNA isoforms, in contrast to the binary representation used in other isoform discovery methods [16, 23, 27, 30]. We still lack the information to model irregular systematic RNA-Seq biases, such as low read coverage in transcript ends and significant read coverage variation unexplained by GC content. But we expect SLIDE to have increased power when such modeling becomes possible with the standardization of RNA-Seq protocols and the improvement of technology. Finally, SLIDE can be easily extended to incorporate mRNA isoform information from EST [33], CAGE [32], and RACE [31] data in addition to RNA-Seq data to refine its linear model and obtain more accurate isoform discovery results.

2.4 Methods

2.4.1 Linear Model Formulation and Identifiability Issue

In the linear modeling of paired-end RNA-Seq data, we first categorize reads into *paired-end bins*. For an n-subexon gene, possible paired-end bins are $\{(i,j,k,l), 1 \le i \le j \le k \le l \le n\}$, whose total number is $m_p = n + 3\binom{n}{2}1_{(n\ge 2)} + 3\binom{n}{3}1_{(n\ge 3)} + \binom{n}{4}1_{(n\ge 4)}$. Then RNA-Seq data is transformed into *bin counts* (i.e., number of reads in each bin), which are further normalized as bin proportions $\mathbf{b}$. Second, we enumerate all the possible isoforms of an n-subexon gene as $I_1, \cdots, I_{2^n-1}$, and denote $\mathbf{p}$ as the isoform proportions to be estimated. Third, we relate unknown $\mathbf{p}$ to observed $\mathbf{b}$ by a design matrix $\mathbf{F}$, where $F_{jk} = \Pr(j\text{th bin} \mid k\text{th isoform})$ (i.e., the conditional probability of observing reads in the jth bin given that the reads are from the kth isoform). (See Subsection 2.4.2 for the modeling of $\mathbf{F}$). Then, we write the following linear model:

$$b_j = \sum_{k=1}^{2^n-1} F_{jk}p_k + \epsilon_j, \quad j = 1, \cdots, m_p, \quad \text{or} \quad \mathbf{b} = \mathbf{F}\mathbf{p} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)$ is the random noise whose components are independent and have mean 0.

We note that the linear model (Equation 2.2) becomes unidentifiable when $m_p < 2^n - 1$ or equivalently $n \geq 9$. The model may also be unidentifiable when $n < 9$ due to possible collinearity of $\mathbf{F}$. To solve this identifiability issue, we reduced the number of parameters $\dim(\mathbf{p})$ by adding a preselection procedure on isoforms. Also, given observed false zero bin counts of certain junction reads, we applied a preselection procedure on observations, too (see Appendix A.1.3). We write the postselection linear model as

$$b_j = \sum_{k=1}^{K} F_{jk} p_k + \epsilon_j, \quad j = 1, \cdots, J. \tag{2.3}$$

We note that the unidentifiability issue still exists in many genes even after the preselection procedures, so sparse estimation is necessary (see Appendix A.1.3).

For single-end data and the combination of both single and paired-end data, we can derive a similar linear model (see Appendix A.1.2).

2.4.2 Modeling of Conditional Probability Matrix

Modeling of the conditional probability matrix $\mathbf{F} = (F_{jk})$, $1 \leq j \leq J$, $1 \leq k \leq K$ is a key part in the estimation of $\mathbf{p}$ (Equation 2.3). In paired-end RNA-Seq data, a mate pair represents ends of a cDNA fragment reversely transcribed from an mRNA transcript. In this sense, F_{jk} is the conditional probability that cDNA fragments with ends in the jth bin are reversely transcribed from mRNA transcripts in the kth isoform. With this interpretation, we model $\mathbf{F}$ with the following three assumptions.

1. The density of a cDNA fragment's starting position (or the density of s_1 in Figure 2.1), denoted by f, is uniform within subexons but proportional to GC content between subexons in an mRNA transcript.

2. The cDNA fragment length ($\ell = e_2 - s_1$ in Figure 2.1) distribution is modeled as truncated Exponential with density denoted by g. This modeling choice is based on empirical observations and Poisson point process approximations (see Appendix A.1.1). SLIDE can also easily take other reasonable fragment length distributions.

3. Starting positions and fragment lengths are assumed to be independent.

In a two-subexon gene example (Figure 2.1), suppose that the two subexons have boundaries $[a_1, b_1]$ and $[a_2, b_2]$. Then, reads in bin $j = (1, 1, 2, 2)$ have $s_1 \in [a_1, b_1 - r + 1]$ and $e_2 \in [a_2 + r - 1, b_2]$. For $k = (1, 2)$, we calculate $F_{jk} = \int_{a_1}^{b_1 - r + 1} f(s_1) \left(\int_{a_2 + r - 1 - s_1}^{b_2 - s_1} g(\ell) \, \mathrm{d}\ell \right) \mathrm{d}s_1$.

For single-end data and the combination of both single and paired-end data, F can be similarly calculated (see Appendix A.1.2).

n	3	4	5	6	7	8	9	10
Datasets 1-2 (37 bp)	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.5
Datasets 3-4 (76 bp)	0.2	0.2	0.2	0.4	0.3	0.4	0.3	0.3
Simulation data (37 bp)	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.5
16 candidate λs:	10^{-6}, 10^{-4}, 10^{-3}, 0.01, 0.04, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1							

Table 2.3: $\lambda^{(n)}$ selection results for different datasets

2.4.3 mRNA Isoform Discovery

In isoform discovery, we expect sparse parameter estimation from the linear model (Equation 2.3), because the number of mRNA isoforms for most *D. melanogaster* genes is below four [29] and far less than the number of possible isoforms K. L_1 penalization approach is widely used for sparse estimation and has applications in high-dimensional and potentially sparse biological data [37]. We also observe that annotated isoforms often contain a large proportion of subexons, and thus expect isoform candidates with more subexons to be more likely true. Hence, we add an L_1 penalty term in the objective function below to limit the number of discovered isoforms as well as to favor the "longer isoforms":

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} \sum_{j=1}^{J} (b_j - \mathbf{F}_j \mathbf{p})^2 + \lambda \sum_{k=1}^{K} \frac{|p_k|}{n_k}, \quad s.t.\ p_k \geq 0, \tag{2.4}$$

where n_k is the number of subexons in the kth isoform and $\mathbf{F}_j$ is the jth row of $\mathbf{F}$. With n_k in the penalty term, p_k would thus be favored if n_k is large. We note that this is a variant of Lasso, a regularization regression method for cases in which the number of parameters to be estimated exceeds the number of observations and most of the parameters are expected to be zeros [34]. The difference between our penalty term and the one in standard Lasso is that the latter only aims to limit the number of discovered isoforms without favoring longer ones. Discussions about choosing L_1 over L_0 regularization and using different likelihoods in the linear model are in Appendix A.

The selection of the regularization parameter λ (Equation 2.4) is by a stability criterion that aims to return the most stable results over different runs of estimation [38]. Because low signal-to-noise ratios in lowly expressed genes may significantly bias the λ selection and genes of the same number of subexons have similar $\dim(\mathbf{p})$ and $\dim(\mathbf{b})$ in Equation 2.4, we group genes by their numbers of subexons n and select an optimal $\lambda^{(n)}$ for each group from 16 candidate values $(\lambda_i)_{i=1}^{16}$ (see Table 2.3). The selection procedure is described in Appendix A, and the chosen $\lambda^{(n)}$ values for datasets 1-4 and the simulation data are in Table 2.3.

R package "penalized" [39] is used in the implementation.

2.4.4 mRNA Isoform Abundance Estimation

The SLIDE linear model (Equation 2.3) can also be used for abundance estimation of discovered or other specified (e.g., annotated) isoform proportions. Because the number of discovered or annotated isoforms is smaller than the number of bin proportions, the linear model is identifiable. Thus, we use nonnegative least squares without a penalty term to estimate the isoform proportions. R package "NNLS" is used in the implementation [40].

2.5 Acknowledgements

I would like to thank Dr. Ci-Ren Jiang, Dr. James B. Brown, Dr. Haiyan Huang and Dr. Peter J. Bickel as co-authors of this work. We would like to thank Qunhua Li and Nathan Boley for their insightful comments during discussions. This work is supported in part by Grants HG004695, HG005639, and EY019094 from the National Institutes of Health.

Chapter 3

Statistical Analysis for Correcting System-Wide Protein Abundance Estimates and Re-Determining Transcriptional Importance in Mammals

3.1 Introduction

The protein products of genes are expressed at very different levels from each other in a mammalian cell. Thousands of genes are not detectably expressed. Of those that are, their proteins are present at levels that differ by five orders of magnitude. Cytoplasmic actin, for example, is expressed at 1.5×10^8 molecules per cell [41], whereas some transcription factors are expressed at only 4×10^3 molecules per cell [42]. There are four major steps that determine differences in protein expression: differences in the rates at which genes are transcribed, mRNAs are degraded, proteins are translated, and proteins are degraded (Figure 3.1). The combined effect of transcription and mRNA degradation together determines mRNA abundances (Figure 3.1). The joint effect of protein translation and protein degradation controls the relative differences between mRNA and protein concentrations (Figure 3.1).

Transcription has long been regarded as a dominant step and is controlled by sequence specific transcription factors that differentially interact with cis-regulatory DNA regions. It has increasingly been realized, however, that the rates of the other three steps vary significantly between genes as well [11, 43, 44, 45, 46, 47, 48]. MicroRNAs, for example, differentially interact with mRNAs of different genes to alter rates of mRNA degradation and protein translation [49, 50, 51, 52, 53].

To quantify the relative importance of each of the four steps, label free mass spectrometry methods have been developed that can measure the absolute number of protein molecules

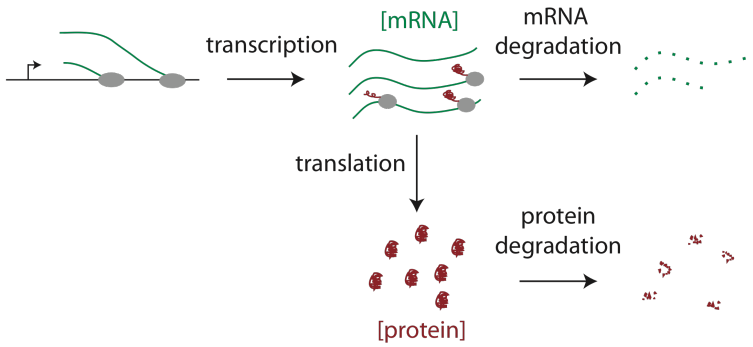

Figure 3.1: The steps regulating protein expression. The steady state abundances of proteins and mRNAs are each determined by their relative rates of production (i.e. transcription or translation) and their rates of degradation.

per cell for thousands of genes [11, 54, 55, 56, 5, 57]. By comparing these data to mRNA abundance data, the relative importance of transcription and mRNA degradation versus protein translation and protein degradation can be determined [11, 56, 57] (Figure 3.1). By measuring mRNA degradation and protein degradation rates as well, the rates of transcription and translation can be additionally infered [11]. Using this approach to study mouse NIH3T3 fibroblasts, Schwanhausser et al. concluded that mRNA levels explain ∼40% of the variability in protein levels and that the cellular abundance of proteins is predominantly controlled at the level of translation [11]. They suggested that transcription is the second largest determinant and that the degradation of mRNAs and proteins play a significant but lesser role.

Our initial reading of the Schwanhausser et al. paper [11], however, suggested that their protein abundance estimates are much lower than established values for individual proteins from the literature. In attempting to characterize the reason for this discrepancy, we also came to suspect that additional sources of experimental error had not been taken into account. Below we describe our re-analysis of this paper and also discuss the relationship between our conclusions and those of Schwanhausser et al. and other system wide studies.

3.2 Results and Discussion

3.2.1 A Non-Linear Underestimation of Protein Abundances

We first noticed that published abundances of 53 mammalian housekeeping proteins [41, 42, 58, 59, 60, 61, 62, 63, 64, 65], 33 of which were derived by SILAC mass spectrometry and 17 by western blot, are on average 16 fold higher than those from Schwanhausser et al.'s label free mass spectrometry data (Dataset S3.1). Once we brought this discrepancy to the authors' attention, they upwardly revised their abundance estimates (see Corrigendum

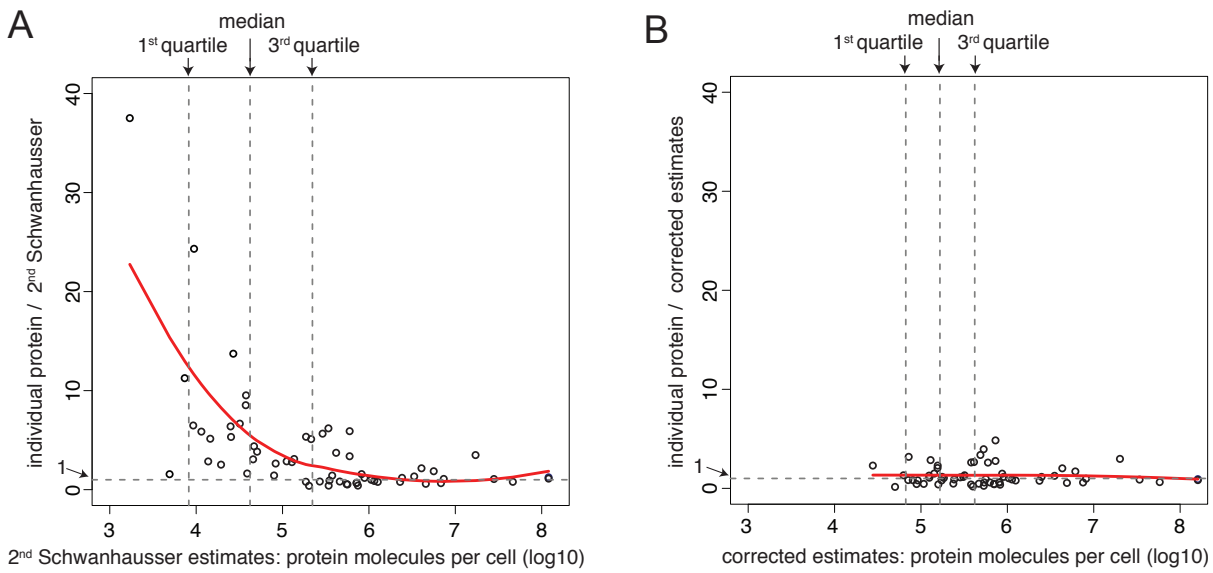

Figure 3.2: A non-linear bias in protein abundance estimates and its correction. (A) The y axis shows the ratios of 61 individually derived protein abundance estimates divided by the abundance estimates from Schwanhausser et al.'s second whole proteome dataset. The x axis shows Schwanhausser et al.'s second whole proteome abundance estimates. The red line indicates the locally weighted line of best fit (Lowess parameter $f = 1.0$), and the vertical dotted grey lines show the locations of the 1st quartile, median and 3rd quartile of the abundance distribution of the 5,028 proteins detected in the whole proteome analysis. (B) The same as panel A except that the whole proteome estimates of Schwanhausser et al. have been corrected using a two-part linear model and the abundances from the 61 individual protein measurements.

[11]). In addition, they provided western blot or Selected Reaction Monitoring (SRM) mass spectrometry measurements for eight polypeptides in NIH3T3 cells. We find, however, that Schwanhausser et al.'s second whole proteome abundances are still lower than the individual measurements for proteins expressed below 106 molecules per cell, with the lowest abundance proteins showing the largest discrepancy (Figure 3.2A; Dataset S3.1).

Western blot and SILAC mass spectrometry measurements show the same discrepancy versus the label free whole proteome data (Dataset S3.1). For example, for proteins expressed below 1 million molecules per cell, the 26 SILAC measurements are a median of 2.95 fold higher than Schwanhausser et al's second estimates, and the 19 western blot measurements are 3.10 fold higher. This suggests that the discrepancy is not due to error in the individual measurements as a similar bias in two independent methods is unlikely.

Of the 61 individual measurements of protein abundance available to us, 15 were made in NIH3T3 cells and 42 were made in HeLa cells. The discrepancy between the second

whole proteome abundances and these individual measurements is not due to differences in expression levels between HeLa and NIH3T3 cells for the following reasons. One, it is unlikely that such a difference would only occur for lower abundance proteins. Two, five of the individual measurements for lower abundance proteins (Orc2, Orc4, HDAC3, NFkB1, and NFkB2) were made in NIH3T3 cells and are on average 3.7 fold higher than the second whole proteome estimates in this same cell line (Dataset S3.1). Three, later in the paper we show that collectively all of the 61 individual proteins measured have on average the same relationship in expression values versus all other cellular proteins in both NIH3T3 and HeLa cells. In addition, as further evidence we note that Schwanhausser et al.'s second estimates for RNA polymerase II and general transcription factors such as TFIIB and TFIIE are only 1.6 fold higher than those in yeast [66] and are 7.1 times less than those in HeLa cells [64]. Yeast cells have 1/40th the volume, 1/200th the amount of DNA and 1/4 the number of genes of NIH3T3 and HeLa cells [67]. Two fold reductions in the concentrations of a single general transcription factor have, in some cases, phenotypic consequence [68, 69, 70, 71]. Thus, it is unlikely that a rapidly dividing mammalian cell could function with much larger reductions in the amounts of all of these essential regulators to levels close to those found in yeast.

3.2.2 Correcting the Non-Linear Bias

Schwanhausser et al. calibrated protein abundances by mixing known amounts of protein standards with a crude protein extract from NIH3T3 cells and then measuring several thousand proteins in the mixture by label free mass spectrometry. The 20 "spiked in" protein standards detected, however, were present at the equivalent $> 8.0 \times 10^5$ molecules per cell, a level that represents only the most highly expressed 11% of the proteins detected (Figure 3.3A) (M. Selbach, personal communication [11]). To convert mass spectrometry signals to protein abundances, Schwanhausser et al. assumed that a linear relationship defined using these 20 "spiked in" standards holds true for proteins at all abundances (Figure 3.3A). The discrepancy between the resulting estimates and individual protein measurements, however, suggests that this assumption is not valid. We therefore employed the 61 individual protein measurements from the literature as they span a much wider abundance range. In a plot of these data vs Schwanhausser et al.'s second whole proteome estimates, we found that a two-part linear regression gave a statistically better fit over a single regression (Figure 3.3B and C) (p-value = 0.002). We then used this two-part regression to derive new abundance estimates for all 5,028 proteins in Schwanhausser et al.'s dataset (Dataset S3.1). As Figure 3.2B shows, the correction removes the non-linear bias.

In our rescaled data, the median abundance protein is present at 170,000 molecules per cell (Figure 3.2B), considerably higher than Schwanhausser et al.'s original estimate of 16,000 molecules per cell and significantly above their second estimate of 50,000 molecules per cell. For low abundance proteins the effect is larger. In our corrected data, the median sequence specific transcription factor is present at 71,000 molecules per cell versus Schwanhausser et al.'s estimates of first 3,500 then 9,300 molecules per cell (Dataset S3.1). Our correction

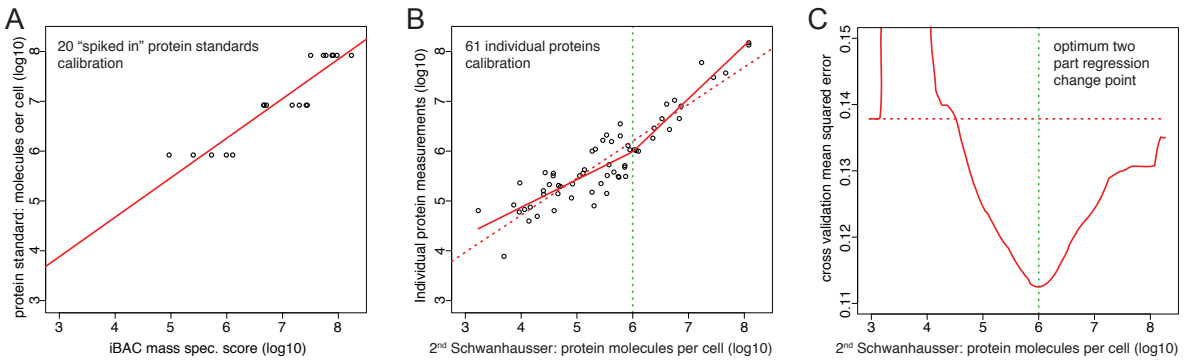

Figure 3.3: Calibrating absolute protein abundances. (A) The relationship between iBAC mass spectrometry signal (x axis) and the amounts of the 20 "spiked in" protein standards (y axis) used by Schwanhausser et al. to calibrate their whole proteome abundances (data kindly provided by Matthias Selbach, Dataset S3.2). The line of best fit is shown (red). (B) The relationship between individually derived estimates for 61 housekeeping proteins (y axis) and Schwanhausser et al.'s second whole proteome estimates (x axis). The two part line of best fit used to correct the second whole proteome estimates is shown (solid red line) as is the single linear regression (dashed red line). (C) The fit of different regression models for the data in panel B. The y axis shows the leave-one-out cross validation root mean square error for each model. The x axis shows the protein abundance used to separate the data for two part linear regressions. The red curve shows the optimum change point for a two part linear model is at an abundance of ~106 molecules per cell. The dashed red horizontal line shows the root mean square error for the single linear regression.

reduces the range of detected abundances by ~50 fold (unlogged) compared to Schwanhausser et al.'s second estimates (Dataset S3.1) and the variance in protein levels from 0.97 to 0.36.

3.2.3 Corrected Protein Abundances Show an Increased Correlation with mRNA Abundances

As an independent check on the accuracy of our corrected abundances, we compared them to Schwanhausser et al.'s RNA-Seq mRNA expression data. Our corrected protein abundances correlate more highly with mRNA abundances than do Schwanhausser et al.'s second whole proteome estimates (compare Figure 3.4A and B). The increase in correlation coefficient is statistically highly significant (p-value $< 10^{-29}$) (see Section 3.4), arguing that our non-linear correction to the whole proteome abundances has increased the accuracy of these estimates. The most dramatic change is that the scatter about the line of best fit is reduced and shows a stronger linear relationship. The 50% prediction band shows that prior to correction the half of proteins whose abundances are best predicted by mRNA levels are expressed over an 11 fold range (unlogged), but after correction they are expressed over a narrower, 4 fold

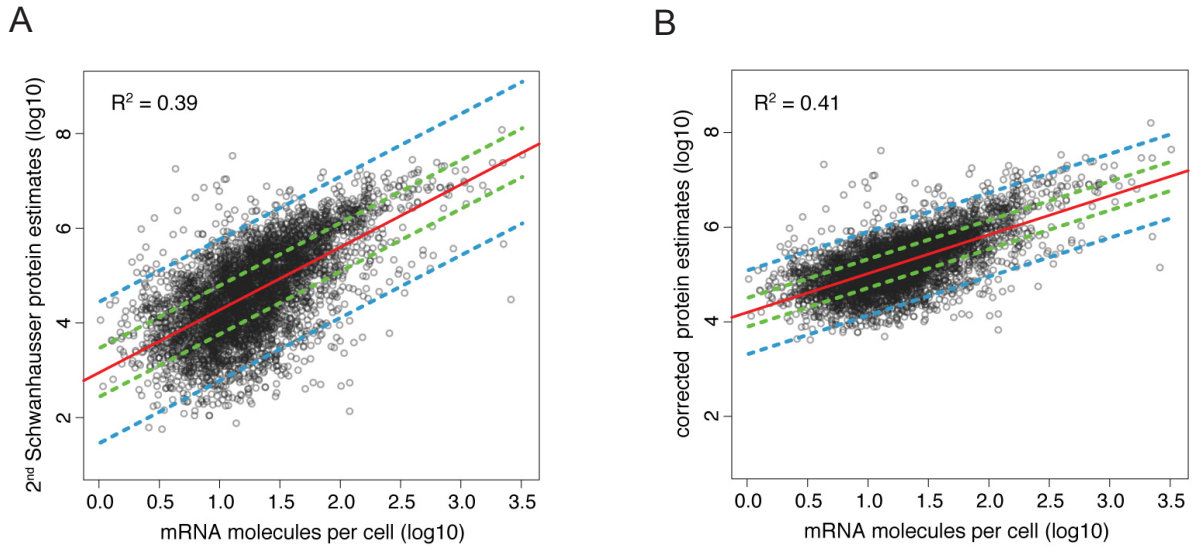

Figure 3.4: Protein abundance estimates versus mRNA abundances. (A) The relationship between Schwanhausser et al.'s second protein abundance estimates vs mRNA levels for 4,212 genes in NIH3T3 cells. The linear regression of the data is shown in red, the 50% prediction band by dashed green lines, and the 95% prediction band by dashed blue lines. (B) The relationship between our corrected estimates of protein abundance vs mRNA levels. The linear regression and prediction bands are labeled as in panel A.

range (Figure 3.4A and B). The correction reduces the width of the 95% prediction band even further, by 18 fold.

For our corrected data, the median number of proteins translated per mRNA is 9,800 compared to Schwanhausser et al.'s original estimate of 900 and their second estimate of 2,800. In yeast, the ratio of protein molecules translated per mRNA is 4,200 - 5,600 [72, 73]. Given that mammalian cells have a higher protein copy number than yeast [67], it is not unreasonable that the ratio in mammalian cells would be the higher.

3.2.4 Estimating the Impact of Molecule Specific Measurement Error

In addition to the above general error in scaling protein abundances, there are additional sources of experimental error that differently affect data for each protein and mRNA. As a result of these molecule specific measurement errors, the coefficient of determination between measured mRNA and measured protein levels—i.e. R^2 shown in Figure 3.4B—is lower than the actual value between true protein and true mRNA levels. With an accurate estimate of the errors, it is possible to calculate the increased correlation expected between true protein

and true mRNA abundances. Because the variance in the residuals in Figure 3.4B (i.e. the displacement along the y axis of data points about the line of best fit) is composed of both experimental error and the genuine differences in the rates of translation and protein degradation between genes, once the experimental error has been estimated, it is also possible to infer the combined true effects of translation and protein degradation.

There are two classes of molecule specific experimental error: stochastic and systematic. Stochastic error, or imprecision, is the variation between replica experiments and is estimated from this variation. Systematic error, or inaccuracy, is the reproducible under or over estimation of each data point, and is estimated by comparing the results obtained with the assay being used to those from gold standard measurements obtained with the most accurate method available.

Schwanhausser et al. limited their estimation of experimental error to stochastic errors. Because our correction of the whole proteome abundances reduces the total variance in measured protein expression levels, we first reestimated the proportion of the variance in the residuals in Figure 3.4B that is due to stochastic measurement error using replica datasets (see Section 3.4). We find that 7% results from stochastic protein error and 0.8% from stochastic mRNA error.

Schwanhausser et al., however, also noted a significant variance between their whole genome RNA-Seq data and NanoString measurements for 79 genes ($R^2 = 0.79$ in Figure S8A in Schwanhausser et al. [11]), though they did not take this into account subsequently. RNA-Seq is well known to suffer reproducible several fold biases in the number of DNA sequence reads obtained for different GC content genomic regions [74, 75]. In contrast, NanoString gives an accurate measure of nucleic acid abundance as correlation coefficients of $R^2 = 0.99$ are obtained when NanoString data are compared to known concentrations of nucleic acid standards [76]. Thus, it is reasonable to consider NanoString as a gold standard that can be used to assess the systematic error in the RNA-seq data by assuming that the variance between the two methods is due mostly to systematic error in RNA-seq. The variance in Schwanhausser et al.'s NanoString/RNA-Seq comparison is equivalent to 23.3% of the variation in the residuals in Figure 3.4B, 29 fold larger than the stochastic component of mRNA error [74, 75].

It is also important to assess the systematic error in the whole proteome abundances as label free mass spectrometry includes such biases [5, 73, 77]. In principle the "spiked in" protein standards in Schwanhausser et al.'s calibration experiment (i.e. the data in Figure 3.3A) should provide gold standard data. In practice, however, the variance in this experiment is significantly higher than that observed between the whole proteome estimates and other abundance data that is known to contain significant error (M. Selbach personal communication). For example, the variance in Schwanhausser et al's calibration experiment would contribute 1.4 fold more to the variance in the residuals in Figure 3.4B than the variance between the corrected whole proteome estimates and the 61 individual protein measurements would. Since no other suitable gold standard is available, we are thus unable to estimate the systematic protein error.

Taking the stochastic protein error as a minimum estimate of protein error and the

variance from the NanoString/RNA-Seq comparison as an estimate of all RNA errors, it can be shown that true mRNA levels explain at least 56% of true protein levels, and by extension protein degradation and translation combined explain no more than 44% (see Section 3.4).

3.2.5 Estimating the Impact of Non-Transcribed Genes

The above estimates, though, only consider the 4,212 genes for which both mRNA and protein abundance data are available. There are many thousands of other genes that are either not detectably transcribed or are more weakly transcribed than these 4,212 genes, and as a result produce little or no protein [78, 79]. To derive a genome wide assessment, therefore, we simulated the true levels of protein expected for an extensive mouse polyA+ mRNA-Seq dataset [78] (see Section 3.4).

Our simulations take into account the trimodal distribution of mRNA expression averaged over a population of animal cells of a single cell type (Figure 3.5) [78, 79]. The 4,212 genes detected by Schwanhausser et al. belong to so-called Highly Expressed (HE) genes, which comprise the most abundant mode and which are expressed above one molecule of mRNA per cell (Figure 3.5). Low Expressed (LE) genes comprise a second mode that are not expressed in the majority of cells butas shown by single molecule fluorescent in situ hybridizationare present at one to several molecules per cell in a small percent of cells. Not Expressed (NE) genes are not detectably expressed in any cells in the population. LE genes tend to be closer to HE genes on the chromosome than are NE genes, and it has been suggested that this proximity may allow escape from repressive chromatin structures in a few cells, explaining the stochastic bursts of rare transcription observed [78, 79].

To account for variation in the expression of individual genes between cells, which all LE genes at a minimum must suffer, our model assumes that the general distribution of mRNA and protein expression levels does not vary from cell to cell even when the expression of individual genes does. For genes in cells that do not express mRNA, an arbitrary, low background level of mRNA expression was chosen because it is not possible to represent zero on a log scale. Conservative values were chosen that are just below the lowest abundances detected in the RNA-Seq dataset. The mRNA expression of each LE gene was divided into a component representing expression of one mRNA molecule in some cells and a second component representing mRNA expression at the arbitrarily defined background level for the remaining cells. This yields 8,763 NE and LE gene equivalents that are not expressed and 12,546 LE and HE gene equivalents that are expressed.

Protein levels for the 12,546 expressed gene equivalents were then simulated using the estimate for the combined variance in translation and protein degradation rates derived previously from the data for 4,212 genes. The 8,763 gene equivalents that express no mRNA are assumed to also express no protein, and thus all such gene equivalents were assigned the same arbitrary, low protein expression value to capture the expectation that there should be no variance in protein expression between them.

For those genes for which Schwanhausser et al. were able to measure both mRNA and protein abundances (i.e. for that particular subset of all HE genes), our model suggests that

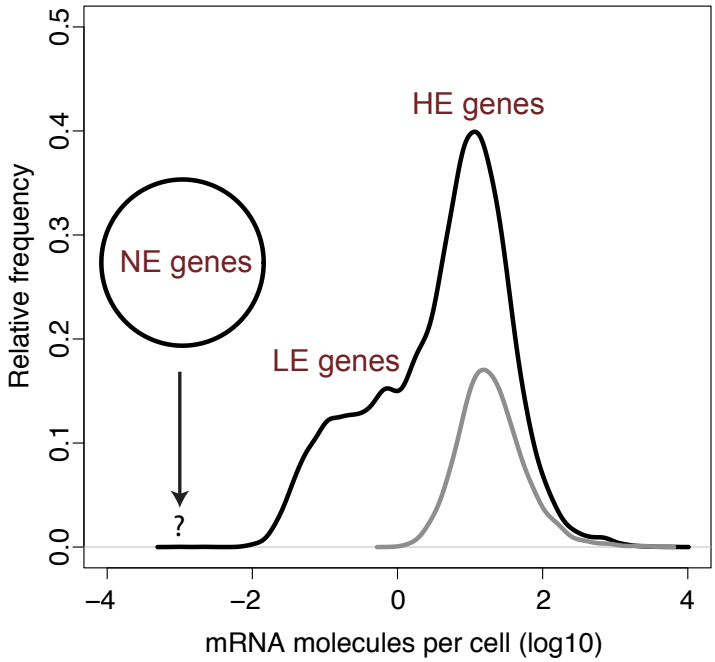

Figure 3.5: The trimodal distribution of mRNA expression levels in animal cells. The black curve shows the frequency distribution for 15,325 genes that give detectable polyA+ mRNA expression in mouse Th2 cells. The two major modes detected for these genes are Highly Expressed (HE) genes centered at 10 molecules of mRNA per cell and Low Expressed (LE) genes centered at 0.1 molecules per cell [76, 77]. The relative frequency of the remaining 5,984 Not Expressed (NE) genes is represented by the area of the circle [76, 77]. The grey curve shows the expression frequency distribution in Th2 cells of the 3,841 genes expessed above 1 molecule per cell that are from the set of the 4,212 genes whose mRNA and protein abundances were detected by Schwanhausser et al. All data has been scaled as described in Section 3.4 and Figure 3.11.

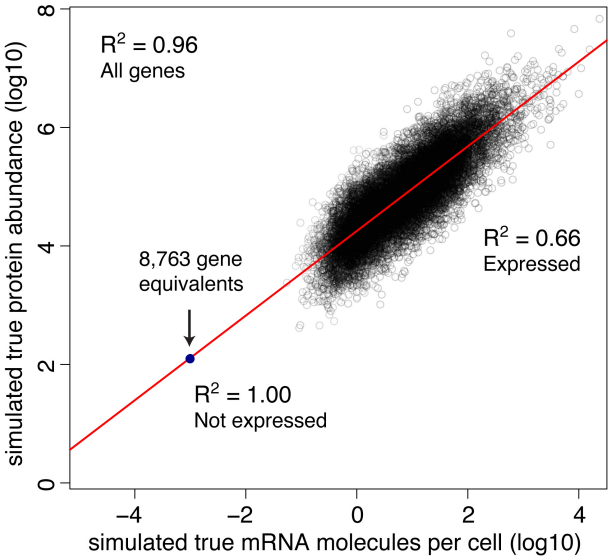

Figure 3.6: Model for true protein abundances versus true mRNA abundances for all 21,309 mouse protein coding genes in Th2 cells. The plots show the result of a typical simulation. The model simulates mRNA and protein expression in each cell of the population by dividing each LE gene into a component expressed at one molecule per cell and a second component expressed at the background level. In addition, the model assumes that genes that are not expressed in a given cell all expresses the identical arbitrary low level of mRNA and protein (arrowed). Results for the 12,546 HE and LE gene equivalents expressed above one molecule per cell (black) and for the 8,763 LE and NE gene equivalents expressed at the background level (blue) are shown. The theoretical R^2 value for all data is 0.96, and for expressed and non-expressed genes separately are 0.66 and 1.0 respectively.

true mRNA levels predict 56% of true protein abundances, the same result obtained for the 4,212 genes in NIH3T3 cells. This indicates that our simulation is quite reasonable. For all 21,309 genes, the R^2 value obtained from the model is 0.96 (Figure 3.6; Table 3.1). We do not believe, however, that the relationship between protein abundance and mRNA across all genes can be summarized by a single R^2 value. The simplest argument is that R^2 is a measure of prediction. The higher the proportion of variance of expressed protein explained by mRNA variance the easier it is to predict expression of a single gene given its mRNA. But predicting a non-expressed gene from its mRNA is trivial. To lump such genes together with expressed genes where prediction is harder seems uninformative and misleading. Instead, we feel it is more appropriate to consider the relationships for expressed and non-expressed genes separately. For the 8,763 gene equivalents that are not expressed, our model suggests that true mRNA levels predict 100% of true protein abundances, and for the 12,546 that are expressed that true mRNA levels predict 65% of true protein abundances.

The higher correlation among the 12,546 expressed gene equivalents compared to that

Table 3.1: The contribution of different steps in gene expression to the variance in protein abundances between genes

	variance in true protein levels ($\log_{10}$)[a]	Percent contribution to variance in true protein levels				
		mRNA	Transcription	RNA degradation	Translation	Protein degradation
Schwanhausser 2nd data 4,212 detected genes[b]	0.97	40%	34%	6%	55%	5%
Measured error strategy 4,212 detected genes[c]	0.34	56%	38%	18%	30%	14%
Measured error strategy 12,546 detected genes[d]	0.43	65%	51%	14%	24%	11%
8,763 non-expressed genes[e]	0	100%	NA	NA	NA	NA
Measured error strategy 4,212 detected genes[f]	0.66	75%	66%	9%	18%	7%
Measured error strategy 12,546 detected genes[g]	0.90	82%	75%	7%	13%	5%

[a]In this column, the value given for Schwanhausser et al.s 2nd data is the variance in their measured protein abundances; the remaining values are our estimate for the variance in true protein levels for different scenarios.

[b]Estimates from Schwanhausser et al. based on the 4,212 genes for which NIH3T3 cell protein and mRNA abundance data are available.

[c]Our estimates for same the 4,212 genes studied by Schwanhausser et al. after correcting the overall scaling of the NIH3T3 cell protein abundance data and taking molecule specific stochastic and systematic experimental error into account.

[d]Our estimates for the model shown in Figure 3.6 for the 12,546 expressed HE and LE gene equivalents in mouse Th2 cells. Protein expression values were modeled using the variance in protein degradation rates measured by Schwanhausser et al and the variance in translation rates estimated in the row above.

[e]Our estimates for the model shown in Figure 3.6 for the 8,763 non-expressed NE and LE gene equivalents in mouse Th2 cells.

[f]Our estimates for same the 4,212 genes studied by Schwanhausser et al. derived using measured translation rates from Ingolia et al.

[g]Our estimates for the 12,546 expressed HE and LE gene equivalents in mouse Th2 cells using protein abundances modeled from the measured variance in translation rates of Ingolia et al and the measured variance in protein degradation rates determined by Schwanhausser et al.

for the 4,212 genes for which data is available ($R^2 = 0.65$ vs 0.56) is due to the fact that the latter set is biased towards more highly expressed genes (Figure 3.5). The addition of many low protein and mRNA expression values will increase the correlation given the assumptions we have made because the variance in protein expression levels increases while the variance in translation and protein turnover rates does not (Table 3.1, column 2). The only circumstance in which consideration of genes expressed at lower levels would not lead to an increase in R^2 would be if the variation in their translation and protein degradation rates were larger than that for the 4,212 detected genes.

3.2.6 Estimating the Relative Importance of Transcription, mRNA Degradation, Translation and Protein Degradation

In addition to determining protein and mRNA abundances, Schwanhausser et al. also directly measured mRNA and protein degradation rates and calculated the percentage that each contributed to the variance in protein abundances. Using this information, it is possible to determine the relative importance of transcription, RNA degradation, translation and protein degradation for different scenarios (see Table 3.1 and Section 3.4). For the 12,546 expressed genes, transcription explains ∼52% of the variance in true protein levels, RNA degradation explains ∼14%, translation ∼24%, and protein degradation ∼10% (Table 3.1). For the 8,763 non-expressed genes, we assume that the absence of transcription is overwhelmingly the reason for the absence of protein expression. Clearly these estimates are tentative and depend on the particular assumptions we have made. We believe, though, that they will prove more accurate than Schwanhausser et al.'s suggestion that translation is the predominant determinant of protein expression and that mRNA levels explain around 40% of the variability in protein levels1 (Table 3.1).

3.2.7 Direct Measurements of Translation Rates Support Our Analysis

Direct measurements of system wide translation rates by Ingolia et al. using ribosome profiling [43] provide independent evidence that translation rates vary less than Schwanhausser et al. suggest. For 95% of the genes whose mRNA was detected, measured translation rates vary only nine fold in mouse embryonic stem cells (Figure 3.7). In contrast, Schwanhausser et al. inferred that for 95% of detected genes' translation rates vary 110 fold (Figure 3.7). Similarly, the variance in translation rates measured by Ingolia et al. is 4.6 fold less than the variance in rates inferred indirectly by Schwanhausser et al. in their model.

Having direct measurements of the variance in translation rates opens up a second strategy to estimate the relative importance of each step in gene expression (Section 3.4). In our first strategy, protein degradation rates and errors in protein and mRNA abundances were determined from direct experimental data; and the variance in true protein levels explained by translation was inferred as that part of the variance in the residuals in Figure 3.4B

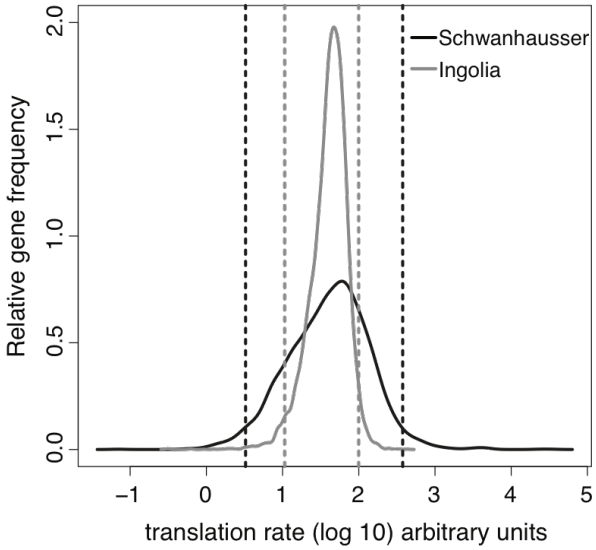

Figure 3.7: Measured versus inferred translation rates. The relative density of ribosomes per mRNA for each gene directly measured by Ingolia et al. [43] (grey lines) compared to the translation rates for each gene inferred by Schwanhausser et al. [11] (black lines). The distribution of values from Ingolia was scaled proportionally to have the same median as that of the Schwanhausser et al. values, and the gene frequencies of the two distributions were normalized to have the same total. The locations of the 2.5 and 97.5 percentiles of each distribution are shown as dashed lines.

that is not explained by the three experimentally measured terms. In our second strategy, translation rates, protein degradation rates and mRNA errors are determined from direct experimental data; and the variance in measured protein levels explained by protein error is inferred as that part of the variance in the residuals in Figure 3.4A that is not explained by the sum of variances of the three experimentally measured components (see Section 3.4). This secondmeasured translationstrategy is thus independent of our rescaling of Schwanhausser et al.'s second protein abundance estimates and of our estimate of stochastic protein measurement error.

According to our second strategy, the variance in true protein levels is 67% of the variance in Schwanhausser et al.'s measured abundances; mRNA levels contribute 76% to the variance in protein expression; transcription 67%; RNA degradation 9%; translation 17%; and protein degradation 7% (Table 3.1). If we model protein expression levels for the 12,546 expressed genes using these variances in translation and protein degradation rates, even higher contributions for mRNA levels (82%) and transcription (75%) are predicted (Table 3.1).

Despite the significant differences in the underlying data and assumption used, our two strategies broadly agree (Table 3.1). Both suggest that the variance in Schwanhausser et al.'s second protein abundance estimates is too high. Both suggest that translation contributes

less to protein levels and that transcription contributes more that Schwanhausser et al. claimed. In effect, Ingolia et al.'s measured rates of translation provide independent support for our rescaling of Schwanhausser et al.'s protein abundances and our estimates of stochastic protein error, and visa versa.

Our second strategy, though, does estimate that mRNA levels and transcription explain a higher percent of protein expression than the first (Table 3.1), but this is not entirely unexpected. In our first strategy, we were not able to take account of systematic, molecule specific errors in protein abundances because appropriate control measurements were not available. Thus, this first strategy could well have underestimated error. In contrast, the second approach estimates all types of protein abundance errors in a single term and thus has the potential to be the more accurate if the error in the ribosome profiling and protein degradation data is not too large. The different results obtained by our two strategies may in addition result, though, because that data that is unique to each approach are subject to variability and are from a different cell line.

Ingolia et al. also showed that translation rates change only several fold upon differentiation of embryonic stems cells and, with the exception of the translation machinery, the change affects all expressed genes to a similar degree [43]. Other system wide studies, including a separate analysis by Schwanhausser et al, also suggest that the differential regulation of translation may be limited to modest changes at a subset of genes [11, 48, 52, 53]. This work seems consistent with our analysis and suggests that translation may be used chiefly for fine tuning protein expression levels.

3.2.8 Implication for Other System-Wide Studies

Two other system wide estimates of protein abundance in mammalian cells are, like Schwanhausser et al.'s, lower than ours. These two reports suggest that the median abundance protein detected is present at 8,000 [54] or 9,700 [55] molecules per cell vs our estimate of 170,000 molecules per cell. Since these lower estimates provide less than 1/10th of the number of histones needed to cover the diploid genome with nucleosomes and are lower than published estimates for a wide array of other housekeeping proteins, it is unlikely that they are accurate.

After completion of the remainder of this manuscript, Wisniewski et al. published protein abundance estimates for HeLa cells that are generally higher than ours and spread over a broader range [80] (Figure 3.8A). These new estimates are also 240% higher on average than the set of individual protein measurements from the literature (Dataset S3, Figure 3.8B). Since over 80% of these individual measurements were made for proteins in HeLa cells, Wisniewski et al.'s estimates must be incorrectly scaled. Using our two part linear regression strategy, we therefore corrected Wisniewski et al.'s whole proteome data (see Section 3.4 and Figure 3.9; Dataset S3), bringing the average variation between the whole proteome estimates and individual protein measurements to within 6% of each other (Figure 3.8B; Dataset S3). Interestingly, the correction dramatically increases the similarity between the distributions of protein abundances in HeLa and NIH3T3 cells for all orthologous proteins (Figure 3.8A).

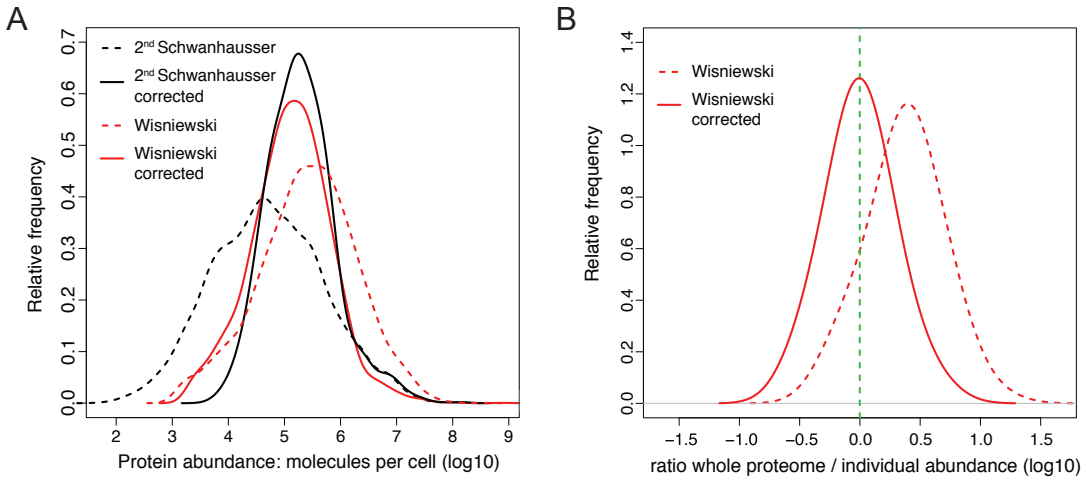

Figure 3.8: Comparison of corrected and uncorrected whole proteome abundance estimates. (A) The distributions of protein abundance estimates for 4,680 orthologous proteins in NIH3T3 cells (black lines) or HeLa cells (red lines). The values from Schwanhausser et al.s second estimates and Wisniewski et al.'s estimates are shown as dashed lines. The values for our corrected abundance estimates are shown as solid lines. (B) The ratios of HeLa cell whole proteome abundance estimates divided by individual measurements from the literature for 66 proteins. Results for the original data from Wisniewski et al. (dashed line) and after these values have been corrected (solid line) are plotted. The green dashed vertical line indicates a ratio of 1.

This establishes the important point, mentioned at the beginning of Section 3.2, that in aggregate the 60+ housekeeping proteins show a similar relationship to the expression values of all other cellular proteins in both cell lines, and thus the discrepancies with the uncorrected whole proteome data are not due to differences in expression levels in HeLa versus NIH3T3 cells. The correction also increases the correlation between HeLa cell protein and HeLa mRNA abundances to a statistically significant extent (p-value $= 6 \times 10^{-20}$) and reduces the 50% and 95% confidence bounds for this relationship by 1.7 fold and 4.6 fold respectively. Wisniewski et al. scaled their protein abundances using the total cellular protein content and the sum of the mass spectrometry signals for all detected polypeptides. They assumed that mass spectrometry signals are proportional to protein abundance. In contrast, our scaling strategy makes no such assumption and instead uses many individual measurements of housekeeping proteins to estimate a multipart (spline) function. The increased correlations obtained with individual protein measurements and with mRNA abundances for two cell lines suggests that our scalings are the more accurate.

Other estimates for the contribution of mRNA levels in determining protein expression in mammals are lower than ours, suggesting that mRNA levels contribute 10%-40% [56, 57]. In comparison, we estimate that mRNA abundance explains 56% - 76% for a set of 4,212

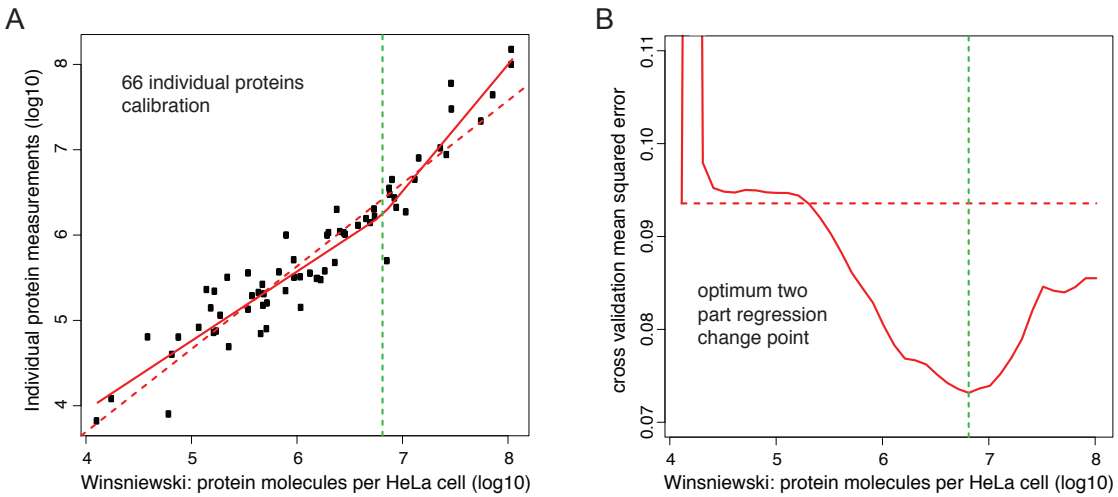

Figure 3.9: Calibrating absolute protein abundances in HeLa cells. (A) The relationship between individually derived estimates for 66 housekeeping proteins (y axis) and Wisniewski et al.s whole proteome estimates from HeLa cells (x axis) (Dataset S3.3). The two part line of best fit used to correct the whole proteome estimates is shown (solid red line) as is the single linear regression (dashed red line). (B) The fit of different regression models for the data in panel A. The y axis shows the leave-one-out cross validation root mean square error for each model. The x axis shows the protein abundance used to separate the data for two part linear regressions. The red curve shows the optimum change point for a two part linear model is at an abundance of $\sim$106.8 molecules per cell. The dashed red horizontal line shows the root mean square error for the single linear regression.

detected proteins, 65% - 82% for all expressed genes and 100% for those genes that are not expressed (Table 3.1). The other groups' studies did not include genes whose protein expression was not detected, and neither took systematic experimental errors into account or made use of direct measures of translation rates. For this reason, we suspect their analyses all underestimate transcriptional importance.

3.3 Conclusions

Quantitative whole proteome analyses can offer profound insights into the control of gene expression and provide baseline parameters for much of systems biology. It is critical, though, to first ensure that these data are correctly scaled, that experimental measurement errors are accounted for as thoroughly as possible, that all genes are considered, and that direct measurements of each step are made. Additional measurements and controls will be needed to derive a more assured system wide understanding of protein and mRNA abundances and the relative importance of each of the four steps in gene expression.

3.4 Materials and Methods

3.4.1 Correcting Protein Abundances

For NIH3T3 cells, all credible individual protein abundance measurements available to us for housekeeping proteins (a total of 61 proteins, Dataset S3.1) were $\log_{10}$ transformed along with the corresponding estimates from Schwanhausser et al.'s second whole proteome dataset. Model selection of different regressive models by leave-one-out cross-validation was used to fit the training data [81]. This showed that a plausible two-part linear regression with a change point at 10^6 molecules per cell (line $< 1 \times 10^6 \ldots$ slope $= 0.56$, intercept $= 2.64$; line $> 1 \times 10^6$ $\ldots$ slope $= 1.06$, intercept $= -0.41$) fit the data far better than by accident (likelihood ratio test bootstrap p-value $= 0.00243$; Figure 3.3B and C). The resulting two-part linear model was used to correct all 5,028 protein abundance estimates (Figure 3.2B, Dataset S3.1).

The null hypothesis that the correlation coefficient of the uncorrected Schwanhausser et al. protein abundance estimates vs mRNA estimates ($R_1 = 0.626$) is equal to that of our corrected protein estimates vs mRNA estimates ($R_2 = 0.642$) was tested using the method for comparing dependent correlation coefficients [82], given that the uncorrected and corrected protein abundance estimates and the mRNA estimates can be assumed to have a multivariate Gaussian distribution. The resulting two-sided p-value $< 10^{-29}$ shows that R_2 is statistically significantly larger than R_1.

To correct protein abundance estimates for HeLa cells [80], the same strategy used for NIH3T3 cells was employed. A two-part linear regression with a change point at 106.8 molecules per cell fit the data far better than by accident (likelihood ratio test bootstrap p-value $= 0.001$) (Figure 3.9). The resulting two-part linear model was used to correct all HeLa cell protein abundance estimates (Figure 3.8; Dataset S3). The correlation of HeLa cell

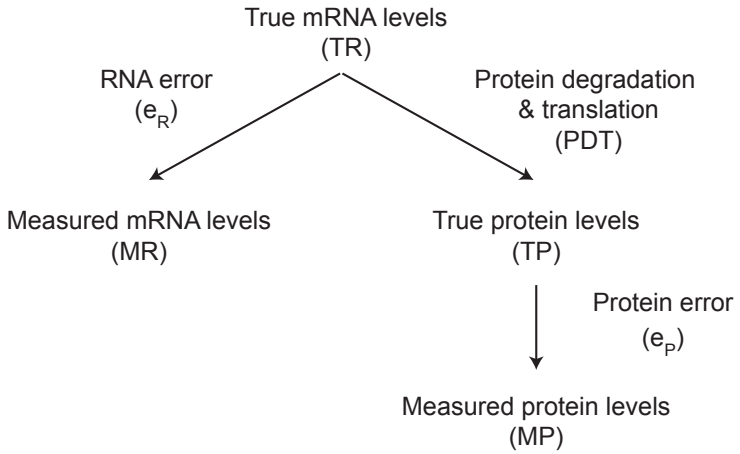

Figure 3.10: The relationship between true and measured protein and mRNA levels.

protein abundance estimates with mRNA abundances was determined using the mean values of replica HeLa cell RNA-Seq datasets from the ENCODE consortium [83] (GEO Accession ID "GSM765402"). The hypothesis that our corrected protein abundances correlate more highly with these HeLa mRNA abundances than the uncorrected estimates was tested as above, resulting in a two sided p-value of 6×10^{-20}.

3.4.2 The Contribution of mRNA to Protein Levels in NIH3T3 Cells: Measured Error Strategy

The variance term in a linear model between measured protein abundance (MP) (response) and measured mRNA levels (MR) (predictor) is decomposed in a standard way (ANOVA [81]) into three components (Figure 3.10). These components of the variance in the residuals represent mRNA measurement error (e_R), protein measurement error (e_P), and the variance in a linear model between true protein abundance (TP) and true mRNA levels (TR) that results from the centered genuine differences in the rates of protein degradation and translation (PDT). The measured protein abundances considered in this case are our rescaled estimates.

Statistically, we can write three linear models from Figure 3.10.

$$TR = b_R MR + c_R + e_R, \tag{3.1}$$

$$TP = bTR + c + PDT, \tag{3.2}$$

$$MP = TP + c_P + e_P, \tag{3.3}$$

where TR, MR, TP, MP are abundance values on a $\log_{10}$ scale; we assume the three sources of variation (e_R, e_P and PDT) are independent random variables with mean 0; the amount

of protein degradation and translation (PDT) is assumed to be independent of true mRNA levels (TR) on the basis of partial evidence: the variance in the residuals in Figure 3.4B is similar for different mRNA abundances; the reversal of the causal relationship between TR and MR in model (3.1) requires another assumption that TR and MR have an approximately joint Gaussian distribution; and finally we assume the slope of TP in model (3.3) can be taken to be 1 because the ratios between the 61 protein published abundance measurements and our corrected estimates are close to 1 (Figure 3.2B). Combining (3.1)-(3.3), we write the linear model between measured protein abundance and measured mRNA levels as

$$MP = bb_R MR + bc_R + c + c_P + be_R + PDT + e_P. \tag{3.4}$$

Based on model (3.4)

1. We first estimated $var(be_R + PDT + e_P)$ as $\hat{\sigma}^2_{all}$ and bb_R as $\hat{b}_{all}$ from fitting the above model with the 8,424 corrected mass spec and RNA-Seq data points pooled from the two replicates (Dataset S3.1). By independence, we have

$$var(be_R + PDT + e_P) = b^2 var(e_R) + var(PDT) + var(e_P).$$

2. We next estimated $var(e_R)$ as $\hat{\sigma}^2_R$ and b_R as $\hat{b}_R$ from fitting model (3.1) with the 77 NanoString ("TR") vs RNA-Seq ("MR") data points, after removing two outliers (Dataset S2).

3. We could not estimate $var(e_P)$ from directly fitting model (3.3), as TP data is not available. As a surrogate, we estimated $var(e_P)$ as $\hat{\sigma}^2_P$ from the following linear model that quantifies the stochastic error in mass spec replicate data:

$$MP_{ij} = \bar{MP}_i + (e_P)_{ij}, j = 1, 2, \tag{3.5}$$

   where MP_{ij} is the corrected mass spec data for the ith protein in the jth replicate in Schwanhausser et al., and $\bar{MP}_i$ is the average of our corrected protein data for the ith protein, $i = 1, \ldots, 4,212$ (Dataset S3.1). Please note that $\hat{\sigma}^2_P$ is potentially an underestimate of the protein error as we only consider the stochastic error, not the systematic error.

4. From the estimates $\hat{\sigma}^2_{all}$, $\hat{b}_{all}$, $\hat{\sigma}^2_R$ and $\hat{\sigma}^2_P$ above, we estimate $var(PDT)$ as

$$\hat{\sigma}^2_{PDT} = \hat{\sigma}^2_{all} - \left(\frac{\hat{b}_{all}}{\hat{b}_R}\right)\hat{\sigma}^2_R - \hat{\sigma}^2_P.$$

Hence, we have successfully decomposed the variance estimate $\hat{\sigma}^2_{all}$, i.e. the estimated variance of residuals between measured protein levels and measured mRNA levels, into 3 components:

- $\hat{\sigma}^2_R$: RNA error (23.3% of $\hat{\sigma}^2_{all}$)
- $\hat{\sigma}^2_P$: protein error (7% of $\hat{\sigma}^2_{all}$)
- $\hat{\sigma}^2_{PDT}$: protein degradation and translation (69.6% of $\hat{\sigma}^2_{all}$)

From the diagram and the above calculation, we also derived the percentage of variability in the unobserved true protein levels explained by the unobserved true mRNA levels.

$$\frac{\hat{\sigma}^2_{MP} - \hat{\sigma}^2_P - \hat{\sigma}^2_{PDT}}{\hat{\sigma}^2_{MP} - \hat{\sigma}^2_P} = 55.9\%,$$

where $\hat{\sigma}^2_{MP}$ is the variance of the corrected measured protein levels.

We separately estimated the stochastic mRNA error from the replicate RNA-Seq measurements of the 4,212 genes (Dataset S3.1). The stochastic mRNA error contributes 0.8% of $\hat{\sigma}^2_{all}$.

3.4.3 The Contribution of mRNA to Protein Levels for All Mouse Genes

To estimate gene expression levels for all genes we employed a deep RNA-Seq dataset that detected polyA+ mRNA for 15,325 protein coding genes in mouse Th2 cells [78]. To place these abundance estimates on the same scale as those of Schwanhausser et al's data, the 3,841 mRNAs expressed above 1 RPKM (reads per kilobase of exon per million mapped reads) in common between the two datasets were identified. The Th2 cell data were then scaled to have the same median and variance for these common genes (Figure 3.11).

To model protein abundances, we first divided each LE gene expressed at less than one molecule of mRNA per cell into two: a fraction of a gene expressed at 1 molecule per cell with a weight w and a fraction of a gene that is not expressed in any cells with a weight $1-w$. The 4,024 LE genes were thus decomposed into 1,245 gene equivalents expressed at 1 molecules per cell and 2,779 gene equivalents that are not expressed. Combining these with the 11,301 HE genes and 5,984 NE genes, we obtained 12,546 HE and LE expressed gene equivalents and 8,763 NE and LE non-expressed gene equivalents. For the measured error strategy, we then simulated the expected levels of protein expressed and true mRNA levels from the 12,546 expressed gene equivalents using $\hat{b}_R$, $\hat{c}_R$, $\hat{\sigma}^2_R$, $\hat{b}$, $\hat{c}$ and $\hat{\sigma}^2_{PDT}$ estimated from our correction to Schwanhausser et al.'s NIH3T3 cell data, see Subsection 3.4.2. The values used to simulate protein levels for the measured translation strategy are described in the next section. For the 8,763 non-expressed gene equivalents, we assigned them true mRNA expression levels of -3.0 ($\log_{10}$) and expected protein expression levels of 2.1 ($\log_{10}$) based on Equation (3.2) estimated previously. Given the weights of the non-expressed and expressed gene equivalents, the weighted coefficient of determination (R^2) was calculated between the simulated expected protein expression levels and true mRNA expression levels for all genes (Figure 3.6). In addition, because we do not view R^2 as an appropriate measure for predicting

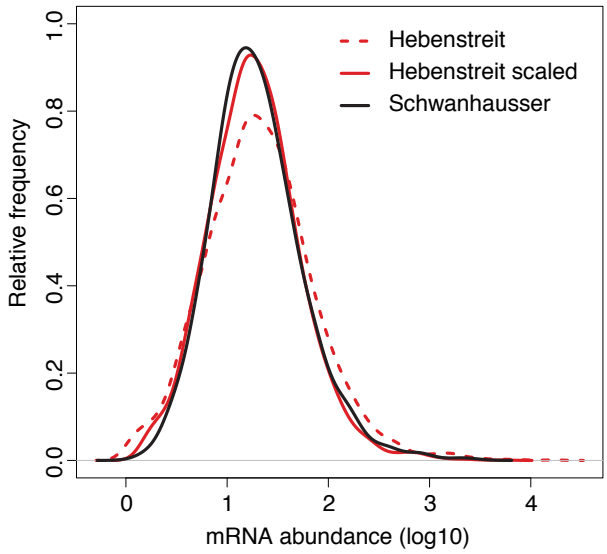

Figure 3.11: Scaling Hebenstreit et al.'s mRNA abundances. The distribution of mRNA abundnaces from three datasets are shown. The 3,841 mRNAs expressed above 1 RPKM in the Hebenstreit et al. RNA-Seq data that are in common with mRNAs detected by Schwanhausser et al were identified (dashed red line). These abundances were then scaled to have the same median and variance as Schwanhausser et al.'s data (solid red line). This scaling was in addition applied to all other genes in the Hebenstreit et al. data and the resulting values used in the simulation shown in Figure 3.6 and in the mRNA expression distribution shown in Figure 3.5.

protein variance for the expressed and non-expressed genes combined (see Section 3.2) we also calculated the R^2 values for the expressed and non-expressed gene equivalents separately (Figure 3.6; Table 3.1).

3.4.4 The Contributions of Transcription, Translation and Protein and mRNA Degradation: Measured Error Strategy

To determine the relative contributions of measured RNA degradation (RD) and measured protein degradation (PD) to the variance in true protein expression (TP), we estimated their variances, $var(RD)$ and $var(PD)$. We took Schwanhausser et al.'s calculated percentages for the contribution of RD and PD to explain the variance of their uncorrected mass whole proteome abundances [11] (6.4% for RD and 4.9% PD, Matthias Selbach personal communication). Since the variance of the 8,424 uncorrected mass spec data points from the two replicates is 0.97, we thus calculated $var(RD)$ and $var(PD)$ as 0.062 and 0.048 respectively. The relative contributions of $var(RD)$ and $var(PD)$ to $var(TP)$ (estimated as $\hat{\sigma}^2_{MP} - \hat{\sigma}^2_P$)

was calculated for several scenarios (Table 3.1). For the same scenarios, we also determined the contribution of transcription to $var(TP)$ as $\frac{var(TR)-var(\text{true } RD)}{var(TP)}$, where $var(TR)$ was estimated as $\hat{\sigma}^2_{MP} - \hat{\sigma}^2_P - \hat{\sigma}^2_{PDT}$, and the contribution of translation as $\frac{var(TP)-var(TR)-var(\text{true } PD)}{var(TP)}$ (Table 3.1).

3.4.5 The Contributions of Each Step of Gene Expression to Protein Levels: Measured Translation Strategy

We calculated the relative contributions of each of the four steps in gene expression by an independent, second approach that does not rely either on our rescaling of Schwanhausser et al.'s protein abundance estimates or on our estimate of stochastic protein errors. Instead, our second approach infers true protein abundance based on Ingolia et al.'s direct measurements of translation rates and on our estimate of RNA measurement error. The measured protein abundances considered are thus Schwanhausser et al.'s second estimates, not our rescaled estimates. A central assumption is that since the variance in Ingolia et al.'s measured translation rates is 4.6 fold less than the variance in the rates of translation inferred by Schwanhausser et al., then the contribution of translation to the variance in true protein levels is 4.6 fold lower than the value provided by Schwanhausser et al.

The variance term in a linear model between measured protein abundance (MP) and measured mRNA levels (MR) was decomposed as before (Figure 3.10) except that the variance in the linear model between true protein abundance (TP) and true mRNA levels (TR) that results from the variance in the rates of protein degradation (PD) and protein translation (PT) were considered separately as cPD and dPT respectively. Similar to our measured error strategy, we can write three linear models using the same assumptions.

$$TR = b_R MR + c_R + e_R, \tag{3.6}$$

$$TP = bTR + cPD + dPT + f, \tag{3.7}$$

$$MP = TP + c_P + e_P, \tag{3.8}$$

Thus, we can write the linear model between measured protein abundance (MP) and measured mRNA levels (MR) for the measured translation strategy as

$$MP = bb_R MR + bc_R + f + c_P + be_R + cPD + dPT + e_P. \tag{3.9}$$

Based on this revised model (3.9)

1. We first estimated $var(be_R + cPD + dPT + e_P)$ as $\hat{\sigma}^2_{all}$ and bb_R as $\hat{b}_{all}$ from fitting the above model with the 8,424 corrected mass spec and RNA-Seq data points pooled from the two replicates (Dataset S3.1). By independence, we have

$$var(be_R + cPD + dPT + e_P) = b^2 var(e_R) + var(cPD) + var(dPT) + var(e_P).$$

2. The estimates of $var(e_R)$ and b_R are the same as those derived previously by our measured error strategy. Thus, we can estimate $\hat{b} = \hat{b}_{all}/\hat{b}_R$.

3. We used the estimate of $var(cPD)$ from Schwanhausser et al., i.e., $0.97 \times 5\% = 0.0475$.

4. From Schwanhausser et al's results, we have $var(dPT) = d^2 var(PT)$ estimated as $0.97 \times 55\% = 0.54$. From Schwanhausser et al.'s estimates for each gene (Dataset S3.1, second tab, column AG) $var(PT)$ has estimate 0.29. Hence, the estimate of d^2 is 1.86. From Ingolia et al, we have a separate, directly measured estimate of $var(PT)$ as 0.06. Using this value to replace that of Schwanhausser et al., we obtained a new estimate of $var(dPT) = d^2 var(PT)$ as $1.86 \times 0.06 = 0.11$.

5. Now we can estimate $var(e_P)$ as $\hat{\sigma}^2_P = \hat{\sigma}^2_{all} - \hat{b}\hat{\sigma}^2_R - \hat{\sigma}^2_{cPD} - \hat{\sigma}^2_{dPT}$ where $\hat{\sigma}^2_{cPD}$ is an estimate of $var(cPD)$ and $\hat{\sigma}^2_{dPT}$ an estimate of $var(dPT)$.

6. Given Schwanhausser et al.'s second 8,424 uncorrected mass spec data, we can also estimate $var(TP)$ as $\hat{\sigma}^2_{TP} = \hat{\sigma}^2_{MP} - \hat{\sigma}^2_P$, where $\hat{\sigma}^2_{MP}$ is an estimate of $var(MP)$.

Given the estimates $\hat{\sigma}^2_{cPD}$ and $\hat{\sigma}^2_{dPT}$ and Schwanhausser et al.'s estimate of the contribution of the variance in RNA degradation (defined as $\hat{\sigma}^2_{gRD}$), we can decompose $\hat{\sigma}^2_{TP}$ as:

- variance explained by PD: $\hat{\sigma}^2_{cPD}/\hat{\sigma}^2_{TP}$
- variance explained by PT: $\hat{\sigma}^2_{dPT}/\hat{\sigma}^2_{TP}$
- variance explained by TR: $1 - \frac{\hat{\sigma}^2_{cPD}}{\hat{\sigma}^2_{TP}} - \frac{\hat{\sigma}^2_{dPT}}{\hat{\sigma}^2_{TP}}$
- variance explained by RD: $\hat{\sigma}^2_{gRD}/\hat{\sigma}^2_{TP}$
- variance explained by transcription: $1 - \frac{\hat{\sigma}^2_{cPD}}{\hat{\sigma}^2_{TP}} - \frac{\hat{\sigma}^2_{dPT}}{\hat{\sigma}^2_{TP}} - \frac{\hat{\sigma}^2_{gRD}}{\hat{\sigma}^2_{TP}}$

Finally, we also determined the expected contributions of each step in gene expression for all 12,546 expressed gene equivalents in mouse Th2 cells. The same procedure described earlier was used except that protein expression levels were simulated using values of $\hat{b}_R$, $\hat{c}_R$, $\hat{\sigma}^2_R$, $\hat{b}$, $\hat{c}$, $\hat{\sigma}^2_{cPD}$ and $\hat{\sigma}^2_{dPT}$ from the measured translation strategy.

3.5 Acknowledgements

I would like to thank Dr. Peter J. Bickel and Dr. Mark Biggin as co-authors of this work. We are indebted to Matthias Selbach for providing his second whole proteome abundance estimates and ancillary data from the Schwanhausser et al. analysis. We acknowledge his patient answering of our questions about the Schwanhausser et al. paper. We also thank Sarah Teichmann for helping us better understand the Hebenstreit et al. analysis of mRNA expression and Susan Celniker, Ben Brown, and David Knowles for constructive comments

on our manuscript. This work was supported in part by NIH grant P01 GM009655. Work at Lawrence Berkeley National Laboratory was conducted under Department of Energy contract DEAC02-05CH11231.

Chapter 4

Comparison of *D. melanogaster* and *C. elegans* Developmental Stages, Tissues and Cells by modENCODE RNA-Seq data

4.1 Introduction

Drosophila melanogaster and *Caenorhabditis elegans*, two mostly intensively studied organisms, serve as model systems for studying molecular, cellular and developmental processes common to higher eukaryotes. Because of their importance and modest genome sizes, *D. melanogaster* and *C. elegans* were among the first organisms with genomes sequenced [84, 85]. The availability of genome sequences and the subsequent microarray technology has enabled molecular studies of *D. melanogaster* and *C. elegans* development on a genome-wide scale. Temporal gene expression patterns have been studied in each organism, suggesting that gene expression changes accompany morphological changes in development [86, 87, 88, 89, 90]. *D. melanogaster* and *C. elegans* are morphologically different and evolutionarily distant organisms, and their developmental life cycles have obvious differences (Figure 4.1A and B): i) *D. melanogaster* has males and females of equal proportions, while wild-type *C. elegans* has 99.5% hermaphrodites and only 0.05% males, ii) *C. elegans* has an alternative developmental path—dauer-interrupted development, a state of developmental arrest that does not exist in the life cycle of *D. melanogaster*. Although conservation in embryonic development in animal species has become a unifying concept since von Baer's observations in the 19th-century [91], little is known about the conservation in post-embryonic development. As a start, comparing genome-wide gene expression patterns throughout developmental stages of *D. melanogaster* and *C. elegans* may help identify unknown conservation in their developmental biology, thus shedding lights on understanding the development of higher species including humans.

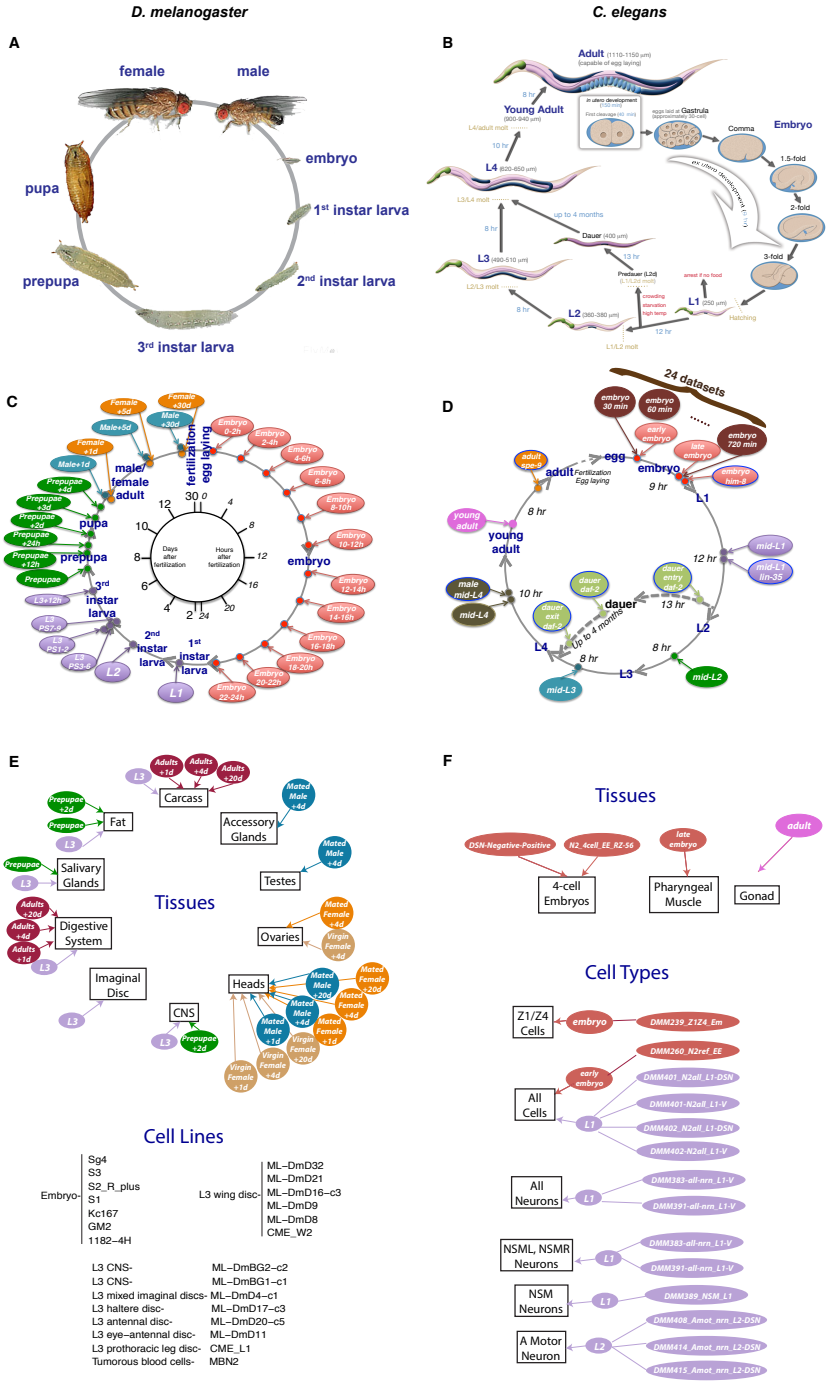

Figure 4.1: Life cycles and modENCODE RNA-Seq datasets of *D. melanogaster* and *C. elegans*. Life cycles of (A) *D. melanogaster* (reprinted with permission from FlyMove by C. Klämbt) and (B) *C. elegans* (reprinted with permission from Wormatlas by D.H. Hall and Z. Altun). modENCODE RNA-Seq datasets of different (C) *D. melanogaster* developmental stages, (D) *C. elegans* developmental stages, (E) *D. melanogaster* tissues and cell lines, and (F) *C. elegans* tissues and cells.

Tissue and cell differentiation is another important topic that has been widely studied in *D. melanogaster* and *C. elegans* for years. Groundbreaking findings include the reversion of germ cells into stem cells in *D. melanogaster* ovaries [92] and the identification of several genes that regulate cell differentiation in *D. melanogaster* or *C. elegans* [93, 94, 95, 96]. To increment the understanding of the molecular basis of tissue/cell differentiation in general, it is necessary to understand the similarity/dissimilarity of different tissues/cells within one species and between different species in the transcriptomic level. Given the considerable biological knowledge on tissue/cell differentiation in *D. melanogaster* and *C. elegans*, the two model organisms serve as good purpose for carrying out a transcriptomic comparison of tissues and cells provided with data availability.

The Model Organism Encyclopedia Of DNA Elements (modENCODE) project aims to identify functional elements in *D. melanogaster* and *C. elegans* genomes and has produced abundant high-throughput RNA sequencing (RNA-Seq) data from different developmental stages, tissues and cells (or cell lines) of the two organisms [53, 54]. The modENCODE RNA-Seq data constitute a good resource for studying genome-wide expression patterns across stages and tissues/cells in the two organisms.

Here we employed the modENCODE RNA-Seq data to compare the developmental stages, tissues and cells of *D. melanogaster* (fly) and *C. elegans* (worm) in terms of genome-wide protein-coding gene expression. First, within each species, we attempted to align developmental stages, tissues and cells by checking the similarity of their associated genes (i.e., genes highly expressed in one stage/tissue/cell but not always highly expressed in all stages/tissues/cells). The within-species alignment results agree with existing knowledge and previously findings, and thus justify the validity of our alignment approach. Next, we aligned developmental stages, tissues and cells between fly and worm by using orthologous genes to link the two species and checking the orthology in stage/tissue/cell-associated genes. Our results provide—for the first time to our knowledge—a comprehensive map between *D. melanogaster* and *C. elegans* developmental stages, tissues and cells, indicating that some conservation exists in the development and tissue/cell differentiation of the two model organisms.

4.2 Results

4.2.1 Identification of Associated Genes for *D. melanogaster* and *C. elegans* Stages / Tissues and Cells (or Cell Lines)

Our goal is to find correspondence, if any, between the developmental stages, tissues and cells (or cell lines) of *D. melanogaster* and *C. elegans* in terms of genome-wide gene expression at the transcriptional level. For every developmental stage, we considered its gene expression characteristics as encoded in "stage-associated genes": the genes highly expressed at that stage but not always highly expressed across all stages. For every tissue and cell (or cell line), we similarly defined "tissue/cell-associated genes" as the genes highly ex-

Table 4.1: Summary of *D. melanogaster* and *C. elegans* genes

# *D. melanogaster* protein-coding genes[b]		# *C. elegans* protein-coding genes[c]		# orthologous gene pairs[a]
all	with worm orthologs	all	with fly orthologs	
13,781	5,467	20,389	5,739	11,403

[a]from modENCODE prediction of fly-worm orthologs (`http://compbio.mit.edu/modencode/orthologs/modencode-orths-2012-01-30/ensembl-v65/modencode.merged.orth.txt.gz`)
[b]genome assembly: BDGP 5.64 [97] (Ensembl assembly 66 [98])
[c]genome assembly: WS 220 [99] (Ensembl assembly 66)

pressed in a particular tissue or cell but not always highly expressed in all tissues and cells. These stage/tissue/cell-associated genes capture signature changes specific to each stage/tissue/cell, which are crucial for understanding gene expression dynamics in development and differentiation. Hence, such genes constitute a basis for aligning the stages, tissues and cells within *D. melanogaster* or *C. elegans*. For between-species alignment, since genes of the two organisms are not directly comparable by using synteny, we focused on their orthologous genes—genes in different species but originated from a single gene of their last common ancestor—and restricted ourselves to the stage/tissue/cell-associated genes having orthologs in the other species.

In this study, we used *D. melanogaster* and *C. elegans* gene annotations from Ensembl [98] and orthologous genes from modENCODE (Table 4.1 and Dataset S4.1). Cufflinks was used to estimate gene expression at different developmental stages or in different tissues/cells (or cell lines) from modENCODE RNA-Seq data (Figure 4.1C-F; Dataset S4.2). The expression estimates are in FPKM (fragments per kilobase of transcript per million mapped reads) units. To identify stage-associated genes, we first normalized every gene's expression profile into z-scores (FPKMs with the mean FPKM over all stages subtracted and then divided by the standard deviation over all stages); then we defined a stage's associated-genes as those with z-scores greater than 1.5 and FPKMs greater than 1 at that stage. Henceforth, those selected stage-associated genes would have a relatively high expression level at that particular stage with respect to a few other stages in the time course and also an absolute expression level above a certain threshold at that stage. We similarly defined tissue/cell-associated genes as the genes with z-scores (FPKMs with the mean FPKM over all tissues and cells subtracted and then divided by the standard deviation over all tissues and cells) greater than 1.5 and FPKM greater than 1 in that particular tissue/cell. Figure 2 provides a summary of the numbers of stage-associated and tissue/cell-associated genes for all the *D. melanogaster* and *C. elegans* developmental stages and tissues/cells. The numbers of genes associated with every stage or tissue/cell range from ∼300 to ∼4,500, where the stages and tissues/cells with higher transcriptional activities, such as early embryonic stages and genital

glands, have larger numbers of associated genes (Figure 4.2I). After we restrict ourselves to the number of stage/tissue/cell-associated genes with orthologs for the purpose of betwee-species alignments, the numbers are greatly reduced and now range from ∼100 to ∼1,600, but their trends across different stages and tissues have not changed much (Figure 4.2II). Most of the stage-associated genes are associated with 2-4 stages. This phenomenon agrees with a biological fact that gene expression levels change continuously during development [100].

4.2.2 Strategy for Aligning *D. melanogaster* and *C. elegans* Stages, Tissues and Cells (or Cell Lines)

Since the so-defined stage- or tissue/cell-associated genes represent transcriptional events specific to a stage or a tissue/cell, we can align different stages or tissues/cells by the similarity of their associated genes. For stage alignment, we compared any two stages by checking the dependence of their associated genes (i.e., the number of shared associated genes if the two stages are of the same species, or the number of associated genes in orthologous pairs if the two stages are of different species). If two stages exhibit significantly strong dependence by a hypergeometric test (see Section 4.4), we called them an "aligned" stage-pair, which can be interpreted as two stages with similar specific gene expression changes.

We used the same strategy to align tissues/cells (or cell lines), and to further align stages to tissues/cells. If a tissue/cell exhibits significant dependence with another tissue/cell or a stage in terms of their associated genes by a hypergeometric test (see Section 4.4), we called them "aligned".

4.2.3 Alignment of Developmental Stages, Tissues and Cell Lines within *D. melanogaster*

We first applied this strategy to aligning developmental stages, tissues and cell lines within *D. melanogaster*, aiming to use the existing extensive knowledge on fly development to justify the validity of our alignment strategy. Our main findings include i) alignment between adjacent developmental stages and ii) alignment of early embryonic stages, female adult stages, ovary tissues, and cell lines. Both results are supported by known biological facts and previous reports.

In the alignment of developmental stages within *D. melanogaster* (Figure 4.3A), first, we expectedly observed that adjacent stages were aligned to each other, a reasonable finding as gene expressions change continuously over time during the development [100].

Second, we found that the earliest embryonic stage (i.e., embryo 0-2 hours) was aligned with female adult stages (i.e., female adult 5-30 days) that bear oocytes, agreeing with previous findings [101]. To determine whether this alignment was a result of maternal gene expression, we compared three gene categories (maternal genes, maternal/zygotic genes and zygotic genes) defined in Lott et al. to the developmental stages. Lott et al. used strain-

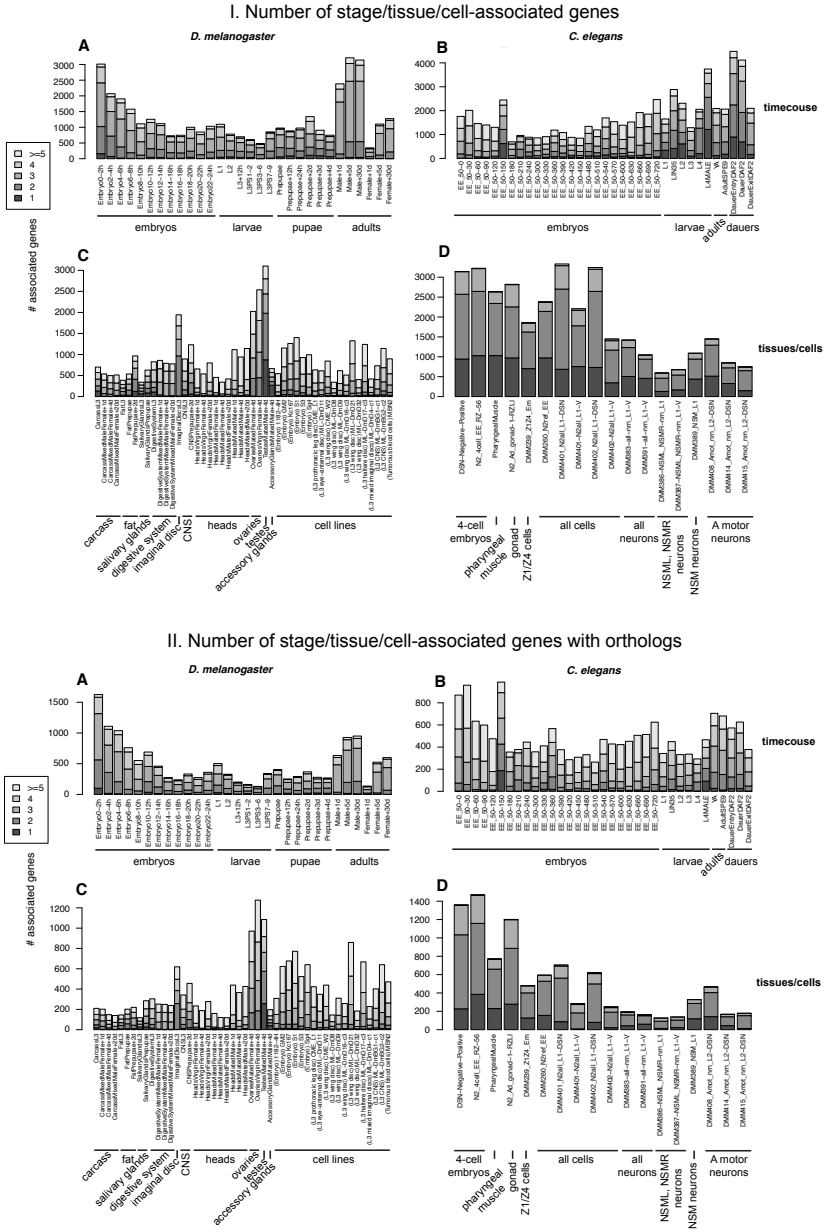

Figure 4.2: Distribution of numbers of stage/tissue/cell-associated genes across different developmental stages, tissues and cells of *D. melanogaster* and *C. elegans*. (I) Number of associated genes (A) across all *D. melanogaster* developmental stages, (B) across all *C. elegans* developmental stages, (C) across all *D. melanogaster* tissues and cell lines, (D) across all *C. elegans* tissues and cells. (II) Number of associated genes that have orthologs in the other species (A) across all *D. melanogaster* developmental stages, (B) across all *C. elegans* developmental stages, (C) across all *D. melanogaster* tissues and cell lines, (D) across all *C. elegans* tissues and cells.

specific time series of *D. melanogaster* gene expression at eight embryonic time points to classify 9,003 genes into three categories: 5,598 maternal genes, 2,210 zygotic genes, and 1,195 maternal/zygotic genes [102]. We used hypergeometric test to test the overlap of the genes in each category and the genes associated with each developmental stage. Figure 3B shows that the maternal and maternal/zygotic genes have significant overlap with the genes associated with early embryonic stages or female adult stages. This result indicates that the observed alignment between fly early embryonic stages and female adult stages is attributable to the expression of maternal and maternal/zygotic genes. Additionally, we observed moderate alignment between fly middle embryonic stages (i.e., embryo 10-16 hours) and larva stages (i.e., L1, L2), and alignment between late embryonic stages (embryo 14-18 hours) and pupa stages (prepupae + 2-3 days) (Figure 4.3A), which are both confirmed by previous findings [86]. These reasonable stage alignment results within *D. melanogaster* justify the validity of our approach as a first check.

In the alignment of tissues and cell lines within *D. melanogaster* (Figur 4.3C), we observed a clear separation of cell lines and tissues after hierarchical clustering (see Section 4.4). A remarkable feature of Figure 4.3C is that the cell lines originated from different tissue sources show a stronger alignment with each other than with the tissues except for ovaries. The observed alignment between ovaries and cell lines is supported by previously reported similarity of cell lines and early embryos [101] and our stage alignment of early embryos and female adults. Figure 4.3C also shows that the head tissues of different fly adults (mated male, mated female and virgin female adults + 1, 4 and 20 days) are aligned with each other, and so are the digestive system tissues of mixed adults at different time points (adults + 1, 4 and 20 days). Such results reveal unexpected stability of gene transcription in the same type of adult tissues across different sex and ages.

In the alignment of *D. melanogaster* tissues/cells to developmental stages (Figure 4.3D), we first observed a clear alignment of the two ovary tissues to early embryonic and female adult stages, which again confirms that maternal genes highly expressed in oocytes are the cause of the alignment we observed between early embryonic and female adult stages (Figure 4.3A). We also found interesting alignment patterns between all the cell lines and early embryonic stages, and between some cell lines and female adult stages. These results are again consistent with our previous alignment results and other reports [101].

4.2.4 Alignment of Developmental Stages, Tissues and Cells within *C. elegans*

We next applied the same alignment strategy to aligning developmental stages, tissues and cells within *C. elegans*, in order to further check the validity of our alignment approach before using it to align stages or tissues/cells between *D. melanogaster* and *C. elegans*. Important findings include i) alignment of early embryonic stages, adult stages, 4-cell embryo tissues, and adult gonad tissues, ii) alignment of tissues/cells from similar organs, and iii) alignment of tissues/cells and their inherent developmental stages.

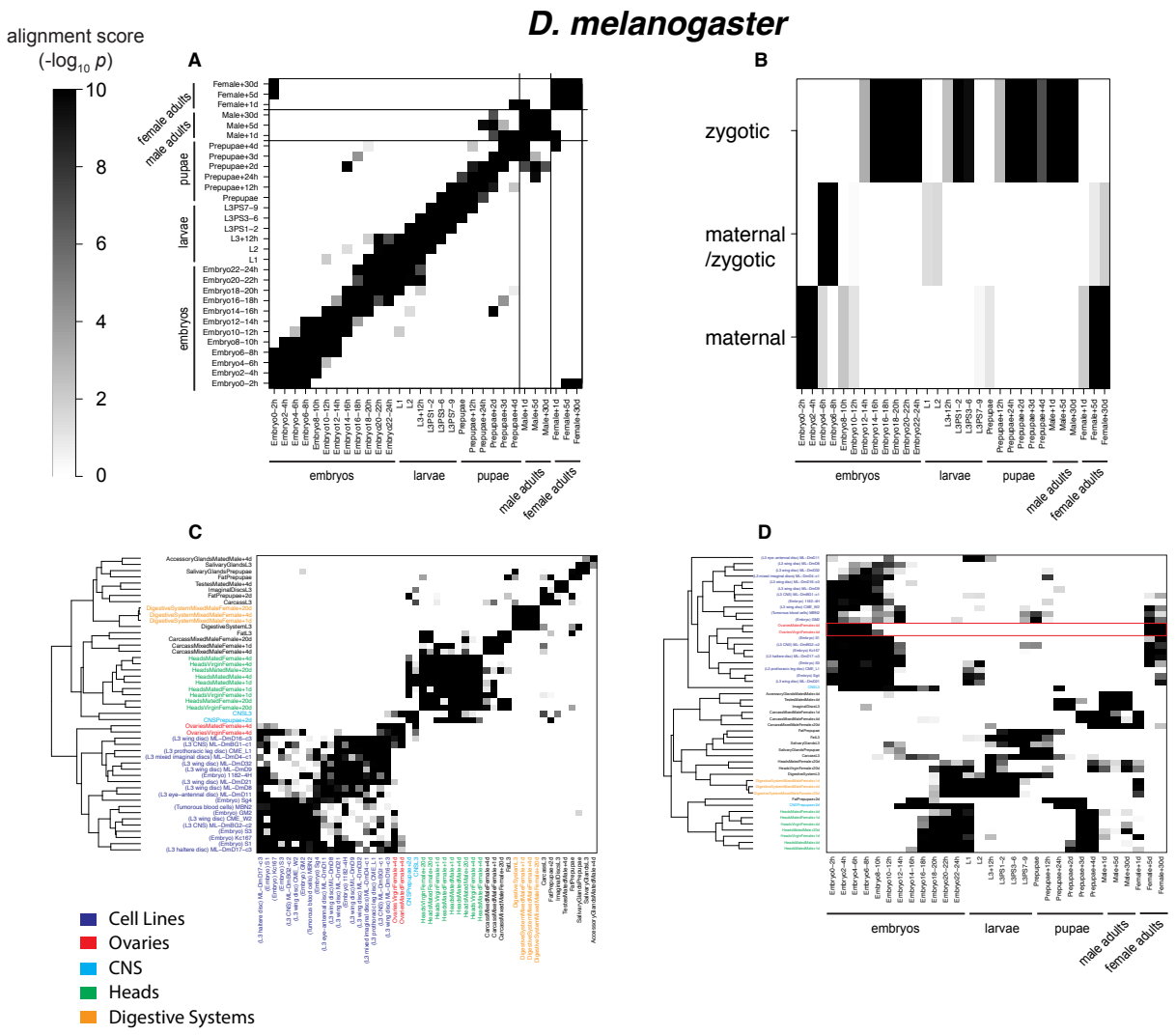

Figure 4.3: Alignment results of different developmental stages, tissues and cell lines within *D. melanogaster*. (A) Stage alignment result. (B) Alignment between three gene categories (maternal, maternal/zygotic, and zygotic) defined by [18] and developmental stages. (C) Tissue/cell line alignment result. (D) Alignment between stages and tissues/cell lines. Hierarchical clustering was applied to order the tissues/cell lines in C and D. Tissues from similar organs are marked with the same color.

In the alignment of developmental stages within *C. elegans* (Figure 4.4A), we observed three interesting alignment patterns: i) alignment of adjacent stages in the timecourse, ii) alignment of early embryonic stages (i.e., early embryo 0-120 minutes) and adult stages (i.e. young adult and adult *spe-9*), and iii) alignment of dauer stages (i.e., dauer entry *daf-2*, dauer *daf-2*, and dauer exit *daf-2*) and larva stages (i.e., L1 and L1 *lin-35*). Similar to the stage alignment within *D. melanogaster*, the observed alignment of adjacent *C. elegans* stages is again a reasonable discovery and supports the validity of our approach. As *C. elegans* are ∼99.5% hermaphrodites that produce all their sperms in the L4 stage and then switch over to producing oocytes [103], the observed alignment of early embryos and adults is attributable to maternal gene expression in worm oocytes. Dauer stages constitute a special developmental path of *C. elegans*, and their alignment with larva stages is consistent with the temporal proximity between dauer and larva stages in the *C. elegans* life cycle (Figure 4.1B).

In the alignment of tissues and cells within *C. elegans* (Figure 4.4B), we found a strong correlation between the alignment and tissue/cell origins after hierarchical clustering on the alignment result: cells extracted from L1-stage worms (all cells) are aligned together; embryonic tissues/cells are aligned with each other. We also observed that 4-cell embryos are aligned with adult gonad tissues. This indicates that the reason of alignment between early embryonic and adult stages is the gene expression in gonad tissues.

In the alignment between developmental stages and tissues/cells within *C. elegans* (Figure 4.4C), we observed three interesting alignment patterns: i) alignment between embryonic tissues/cells and early embryonic stages, ii) alignment between 4-cell embryos/adult gonad tissues and early embryonic/adult stages, and iii) alignment between cells extracted from L1-stage worms (all cells and neurons) and late embryonic to larva developmental stages. These patterns show a strong correlation between the tissues/cells and their inherent developmental stages, a phenomenon not observed in the stages vs. tissues/cell lines alignment within *D. melanogaster*.

4.2.5 Mapping of Developmental Stages, Tissues, and Cells (or Cell Lines) between *D. melanogaster* and *C. elegans*

After verifying the reasonableness of our stage/tissue/cell alignment results within *D. melanogaster* and *C. elegans* respectively, we applied our alignment strategy to aligning developmental stages and tissues/cells (or cell lines) between the two species.

As the very first attempt to study the correspondence between the life cycles of *D. melanogaster* and *C. elegans*, we aligned their developmental stages on the basis of shared orthologs in their stage-associated genes. Figure 4.5A shows an interesting and surprising stage alignment result. First, a collinear alignment pattern is observed between fly early embryos through larvae and worm early embryos through larvae. Second, we found another collinear alignment pattern including: fly L1 larvae – worm middle embryos, fly prepupae – worm late embryos, and fly female adults – worm adults. The two parallel collinear patterns indicate a division of the fly life cycle into two parts: the first part (from fly early embryos to

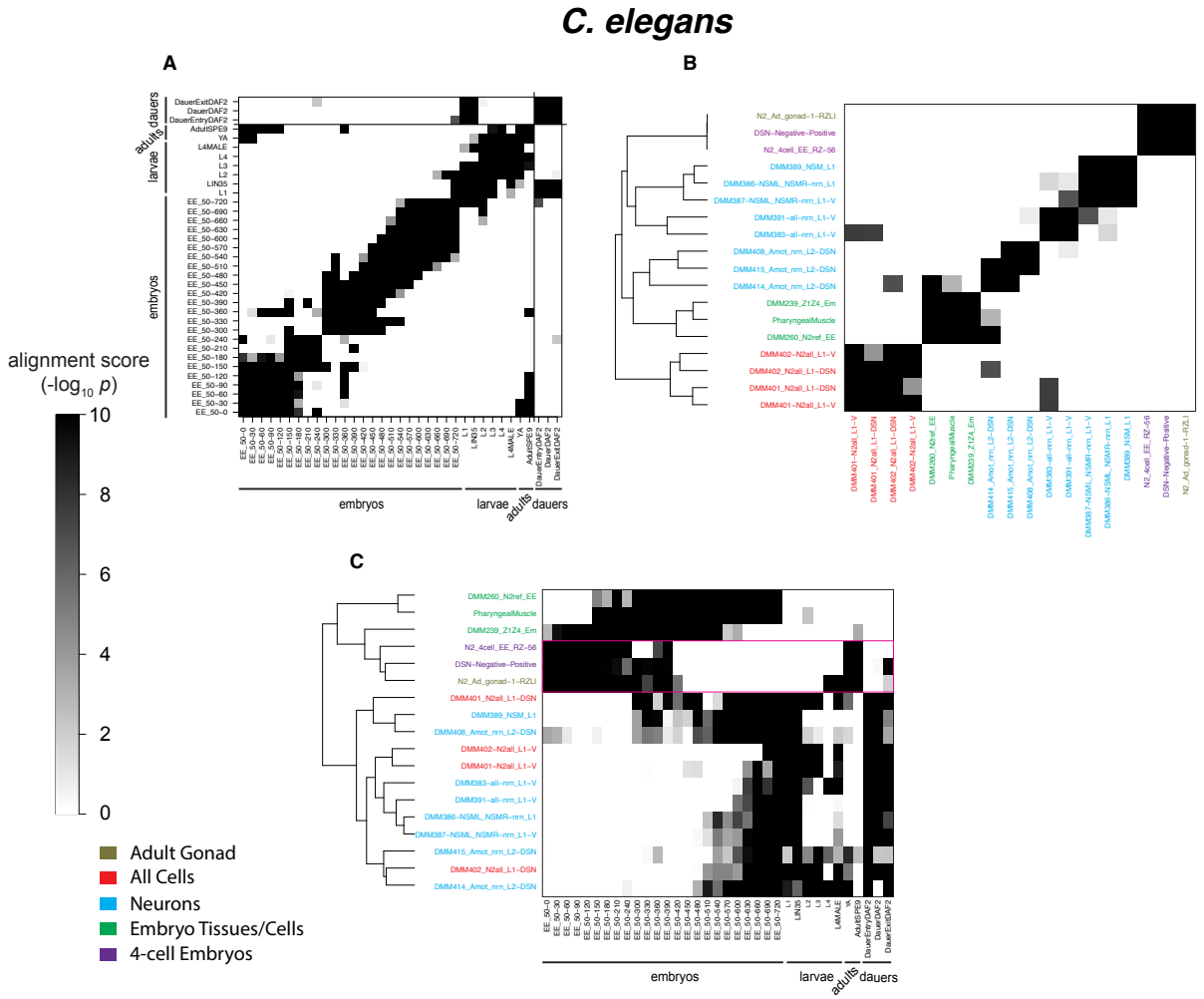

Figure 4.4: Alignment results of different developmental stages, tissues and cell lines within *C. elegans*. (A) Stage alignment result. (B) Tissue/cell alignment result. (C) Alignment between stages and tissues/cells. Hierarchical clustering was applied to order the tissues/cells in B and C. Tissues and cells with similar origins are marked with the same color.

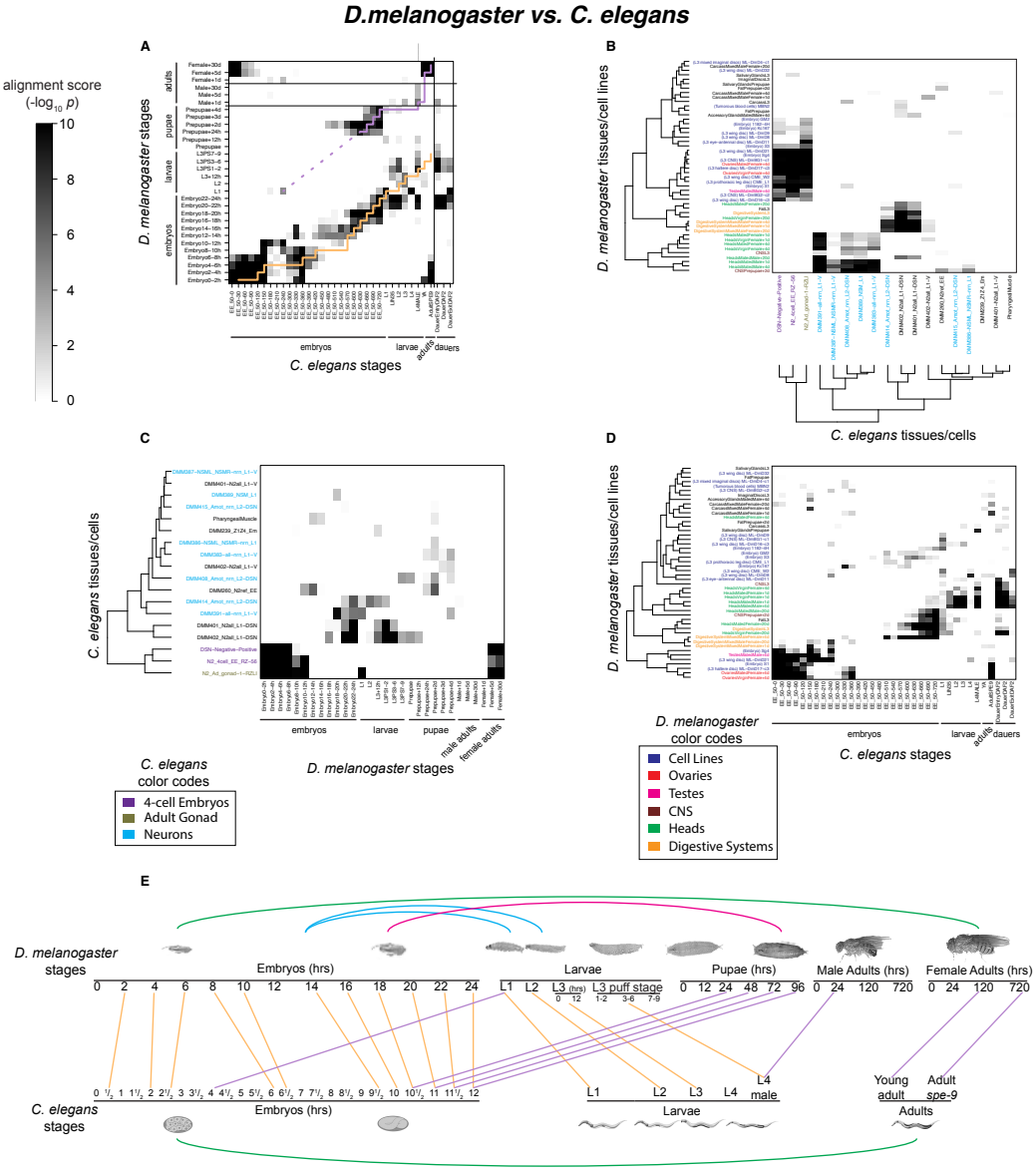

Figure 4.5: Alignment results of different developmental stages, tissues and cell lines between *D. melanogaster* and *C. elegans*. (A) Alignment between *D. melanogaster* and *C. elegans* developmental stages. The orange and purple stepwise lines were found by maximizing the sum of alignment scores of the stage-pairs they pass through, to represent the two parallel collinear stage alignment patterns. (B) Alignment between *D. melanogaster* tissues/cell lines and *C. elegans* tissues/cells. (C) Alignment between *D. melanogaster* developmental stages and *C. elegans* tissues/cells. (D) Alignment between *D. melanogaster* tissues/cell lines and *C. elegans* developmental stages. Hierarchical clustering is applied to order the *D. melanogaster* tissues/cell lines and *C. elegans* tissues/cells in B, C and D. Tissues and cells with similar origins are marked with the same color. (E) A cartoon summary of main alignment results in A.

larvae) is aligned with the worm life cycle except for the worm adult stages, and the second part (from fly prepupae to adults) is aligned with the worm life cycle except for the worm early embryonic stages (Figure 4.5E). Some worm stages, including middle embryos (early embryos + 240 minutes, i.e. "EE_50_240"), late embryos (early embryos + 480-720 minutes, i.e. "EE_50_480"-"EE_50_720"), and L4 male larvae (i.e., "L4male"), are each aligned with two blocks of fly stages ("A", "B" and "C" in Figure 4.6A), denoted as early alignment (alignment between worm stages and the block of earlier fly stages) and "late alignment" (alignment between worm stages and the block of later fly stages).

To figure out the reasons for such two-to-one fly-worm stage alignments, we asked two questions: a) given a worm stage, are its aligned two blocks of fly stages aligned with each other in the stage alignment within *D. melanogaster*? b) are the early alignment and late alignment due to the same set of fly and worm genes? To answer question a), we compared the three stripes of fly stages ("A", "B" and "C" in Figure 4.6A) with the stage alignment result within *D. melanogaster* (Figure 4.3A). The two fly blocks in stripe "A" (middle embryos and L1 larvae) are moderately aligned within *D. melanogaster* (Figure 4.3A), and so are the two fly blocks in stripe "B" (middle to late embryos and pupae). However, there is no clear alignment between the two fly blocks in stripe "C" (late embryos to prepuape and male adults). Hence, the off-diagonal stage alignments in Figure 4.3A cannot sufficiently explain all the two-to-one fly-worm stage alignments in Figure 4.6A, and the answer to question a) is no. To answer question b), we classified the fly and worm stage-associated genes by their involvement in the early and late alignments. For example, suppose that we have a worm stage W aligned with two fly stages FE (the earlier stage) and FL (the later stage), i.e., the early alignment is $W - FE$, and the late alignment is $W - FL$. Also suppose that W has stage-associated genes w_1, w_2, w_3, w_4; FE has stage associated genes f_1, f_2, f_4; FL has stage associated genes f_1, f_3, f_5. Orthologous gene pairs between worm and fly are $w_1 - f_1$, $w_2 - f_2$, $w_3 - f_3$, and $w_4 - f_4/f_5$. Hence, the ortholog pairs that lead to the early alignment is $w_1 - f_1$, $w_2 - f_2$, and $w_4 - f_4$; the ortholog pairs that lead to the late alignment is $w_1 - f_1$, $w_3 - f_3$, and $w_4 - f_5$. This simple example demonstrates that we can classify gene ortholog pairs involved in any two-to-one fly-worm stage alignment into four categories: i) ortholog pairs where both worm and fly genes are only involved in early alignment (e.g. $w_2 - f_2$), ii) ortholog pairs where both worm and fly genes are only involved in "late alignment" (e.g. $w_3 - f_3$), iii) ortholog pairs that lead to both early alignment and late alignment (e.g. $w_1 - f_1$), and iv) ortholog pairs where the same worm gene but different fly genes are involved in "early alignment" and "late alignment" (e.g. $w_4 - f_4/f_5$). We did this classification for each of the stripes "A", "B" and "C" in Figure 4.6A and summarized the results in Figure 4.6C, which shows that orthologs in all the four categories contribute to the observed two-to-one fly-worm stage alignments. Hence, the answer to question b) is also no. This again confirms that within-fly stage alignments lead by the fly genes in category iii) is not the only reason for these two-to-one fly-worm stage alignments, and the contribution of category iii) decreases from stripe "A" to "C". Figure 4.6C also shows that the contribution of category iv), i.e. many-to-one fly-worm orthologs, increases from stripe "A" to "C". Fly genes in category iv) would not lead to within-fly stage alignment, and their roles in these

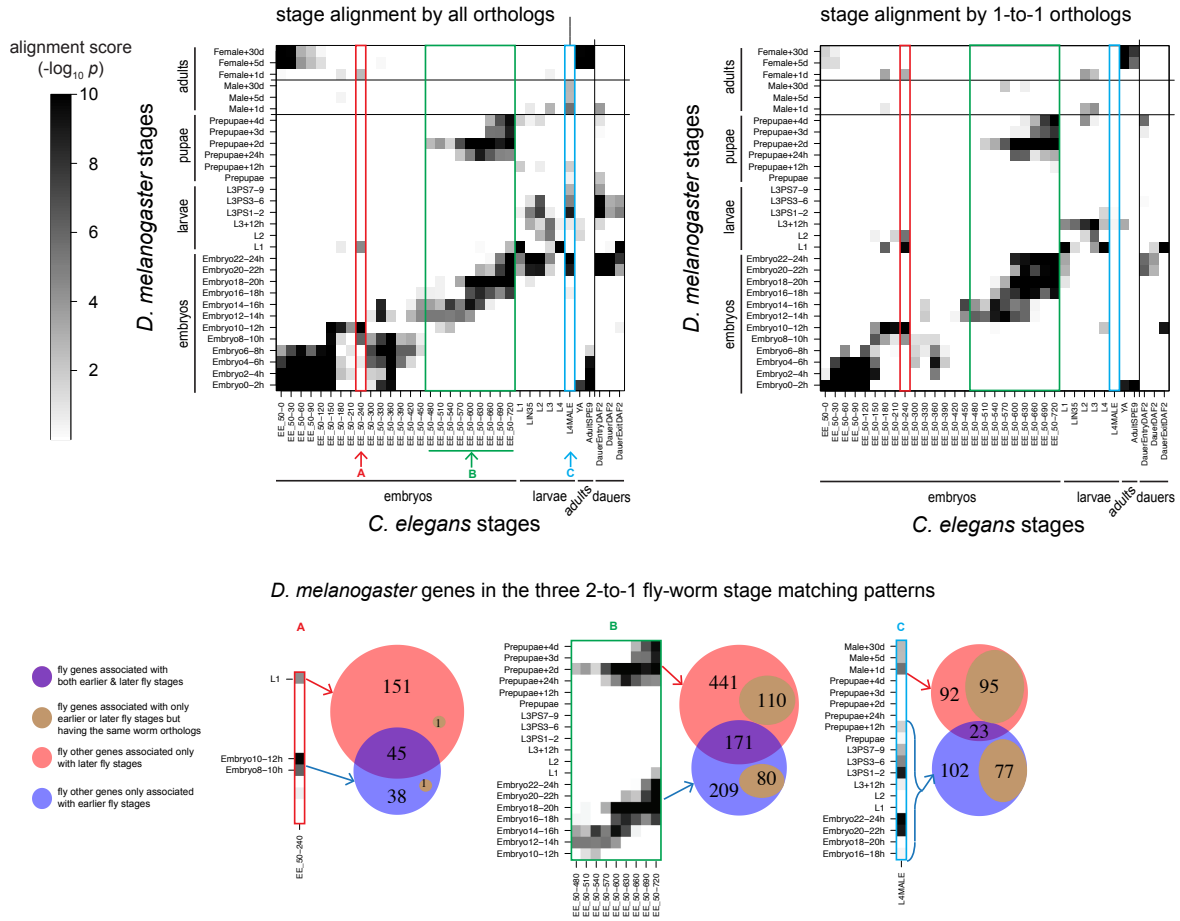

Figure 4.6: Interpretation of the observed two-to-one fly-worm stage alignment patterns. (A) Alignment between *D. melanogaster* and *C. elegans* developmental stages by using all ortholog pairs. The red, green and cyan boxes marked the three main two-to-one fly-worm alignment stripes. (B) Alignment between *D. melanogaster* and *C. elegans* developmental stages by using only one-to-one ortholog pairs. The red, green and cyan boxes are in same positions as the corresponding boxes in (A). (C) Classification of ortholog-pairs based on their involvement in the "early alignment" (the lower block of fly stages) and late alignment (the upper block of fly stages) in each stripe ("A", "B" or "C") in (A).

two-to-one fly-worm stage alignments indicate that different fly orthologs of the same worm gene are highly expressed at different time points in the fly development, implying some type of redundancy in gene orthology. To further check the roles of orthologs in category iv) that lead to these two-to-one fly-worm stage alignments, we re-aligned the fly and worm stages by restricting ourselves to only one-to-one fly-worm orthologs, and the results are shown in Figure 4.6B. By comparing Figures 4.6A and B, we can see that the stage alignment patterns are generally the same but with slight differences. Stripe "C" becomes negligible in Figure 4.6B, because its early and late alignments in Figure 4.6A are largely attributable to the orthologs in category iv), which are removed in Figure 4.6B. On the contrary, stripes "A" and "B" become strengthened in Figure 4.6B, implying that their alignments are mostly lead to by one-to-one fly-worm orthologs and the orthologs in category iv) do not play a key role.

In addition to the two parallel collinear patterns, we also observed alignment between fly early embryos and worm adults, and between fly female adults and worm early embryos (Figure 4.5A). These results, coupled with the alignment between fly female adults and worm adults, indicate strong orthology between maternal genes in the two species.

To summarize, the fly-worm stage alignment results (Figure 4.5A and E and Figure 4.6) are the first findings showing the similarity of worm and fly developmental timecourses in terms of system-wide orthologous gene expression. An unexpected pattern of two parallel collinear alignments is revealed, which is symbolic to a twice repetition of the worm life cycle in the fly early and late life cycle. Both within-fly stage alignment and the many-to-one fly-worm orthologs play important roles in leading to the two parallel collinear patterns.

In order to understand the similarity of tissues/cells and developmental stages between the two species, we used the same between-species alignment approach to align i) worm tisuses/cells with fly stages (Figure 4.5C) and ii) fly tissues/cells with worm stages (Figure 4.5D). Figure 4.5C shows that worm 4-cell embryos and adult gonad tissues are aligned with fly early embryonic and female adult stages. Figure 4.5D shows that fly gonad tissues (ovaries and testes) and several cell lines are aligned well with worm early embryonic and adult stages. These two findings are again results of maternal gene expression, implying again strong orthology between maternal genes in both species.

Finally, we attempted to align tissues and cells (or cell lines) between fly and worm. The alignment result in Figure 4.5B shows two interesting patterns. First, most worm neuron tissues are aligned with fly heads (in adults) and CNS tissues (in larvae and pupae), indicating strong orthology of genes with neural functions in both species. Second, worm 4-cell embryos and adult gonad tissues have a clear alignment with fly cell lines and adult gonad tissues (ovaries and testes). This is again a proof of strong orthology of maternal genes in both species.

4.3 Discussion

We developed a hypothesis testing approach to align developmental stages, tissues and cells (or cell lines) within and between *D. melanogaster* and *C. elegans*, on the basis of their transcriptome-wide protein-coding gene expression. Our approach centers on i) using orthologous genes to link the two species and ii) identifying stage/tissue/cell-associated genes to represent specific transcriptional events for developmental stages, tissues and cells. Our approach differs from a more intuitive approach, that is, to calculate the correlation coefficient (Pearson or Spearman) of gene expression levels in two samples (stages, tissues or cells), which has been widely used in microarray and RNA-Seq analyses. We first tried the correlation approach but found that neither Pearson nor Spearman correlation is a good measure for aligning developmental stages within *D. melanogaster* or *C. elegans* (Figure 4.7). Pearson correlation is not robust to outliers and depends highly on the gene expression estimates (in FPKM units). Spearman correlation is a better measure than Pearson correlation as it is more robust to outliers. However, due to housekeeping genes that are constantly highly expressed across all developmental stages, Spearman correlation still does not lead to clear alignment patterns in neither species. Unlike the correlation approaches, our approach does not use all genes but focuses on small subsets of genes that capture specific transcriptional events in different developmental stages. Genes whose expression levels have little variance across different stages are thought to contain little information on stage alignment and are thus excluded in our approach. Hence, our approach can provide more clearcut alignment results compared to the correlation approaches. Also, our approach is based on the selected subsets of genes, not their absolute expression levels, and is thus more robust to biases and errors in gene expression estimates.

We first applied our approach to aligning developmental stages within *D. melanogaster* and *C. elegans* as a sanity check. Quite reasonably, stages temporally adjacent to each other are aligned to each other in both species. Another reasonable finding is the alignment of early embryos with female adults in fly and with adults in worm, which is a result of maternal gene expression in oocytes. Other unexpected findings, including the alignment of fly middle embryos and larvae and the alignment of fly late embryos and pupae, are supported by previous findings.

After passing the sanity check, we next applied our approach to aligning tissues and cells (or cell lines) within *D. melanogaster* and *C. elegans*. Interesting and reasonable findings include i) alignment of cell lines and ovary tissues in fly, ii) alignment of tissues from the same organ in fly, iii) alignment of early embryonic tissues and gonad tissues in worm, and iv) alignment of tissues from similar developmental stages in worm. We also aligned tissues/cells with developmental stages within each species, and found that early embryonic and adult gonad tissues are aligned with early embryonic and adult (female adult in fly) stages in both species. These reasonable results further justify the validity of our approach.

Given the above within-species alignment results, we finally used our approach to align developmental stages between *D. melanogaster* and *C. elegans*. Surprisingly, our result revealed two parallel alignment patterns between the timecourses of the two species. Both

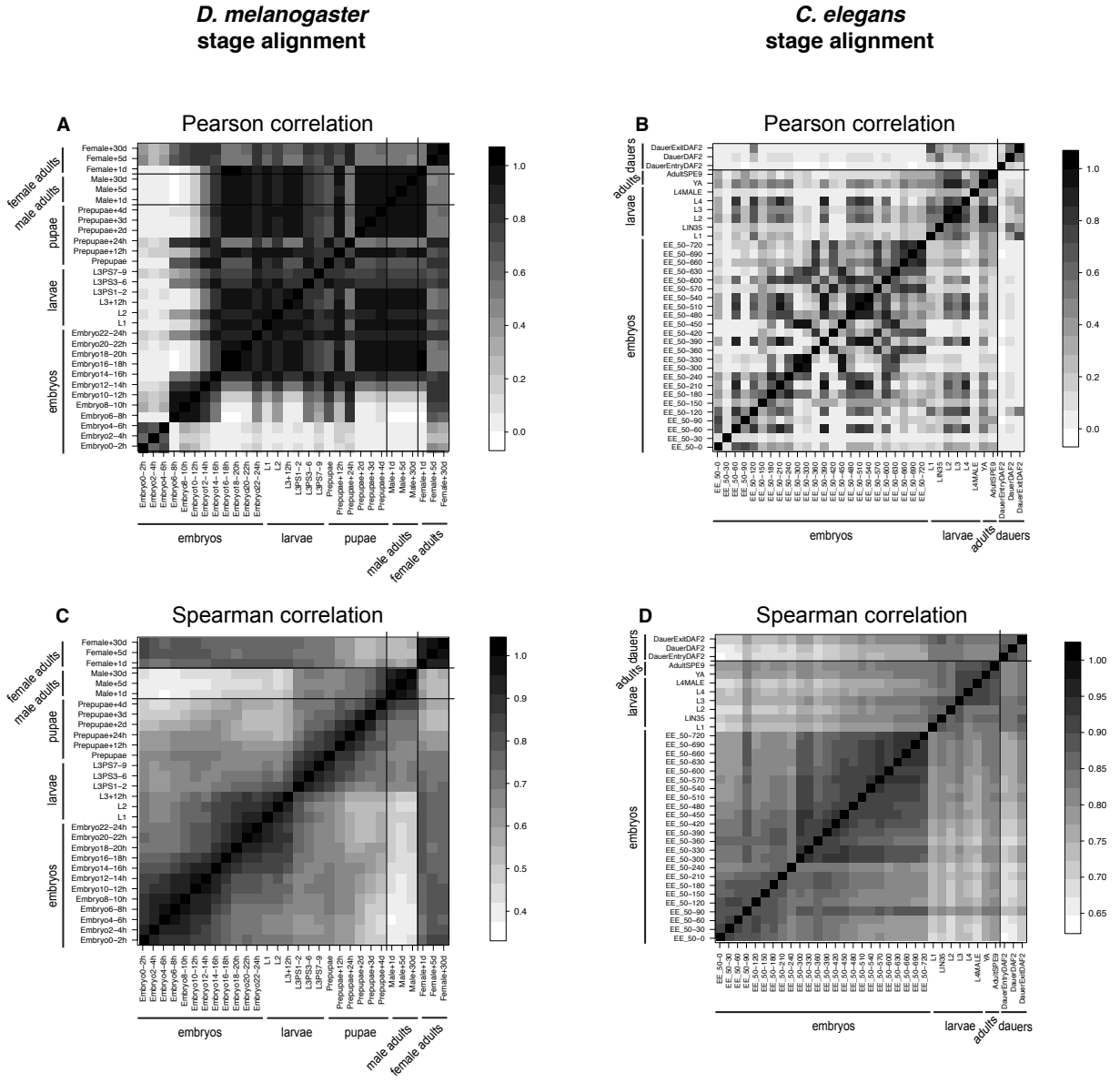

Figure 4.7: Intuitive correlation approaches of aligning developmental stages within *D. melanogaster* and *C. elegans*. (A) Alignment by calculating a Pearson correlation of all protein-coding gene expression levels for every pair of *D. melanogaster* stages. (B) Alignment by calculating a Pearson correlation of all protein-coding gene expression levels for every pair of *C. elegans* stages. (C) Alignment by calculating a Spearman correlation of all protein-coding gene expression levels for every pair of *D. melanogaster* stages. (D) Alignment by calculating a Spearman correlation of all protein-coding gene expression levels for every pair of *C. elegans* stages.

alignment patterns cover most of the worm timecourse but correspond to differnet parts of the fly timecourse. This result can be interestingly interpreted as a recapitulation of the worm life cycle twice in the fly life cycle. In our investigation for a possible explanation, we found that the two parallel alignment patterns have some connections with the alignments between non-adjacent stages within fly. But that is not the only cause for the parallel patterns. We found that gene expression of many-to-one fly-worm orthologs also partially lead to such patterns. This implies the possibility of more redundancy in developmental gene functions in fly than in worm.

At last, we aligned tissues/cells to developmental stages and tissues/cells to tissue/cells between the two species. Interesting findings include i) alignment of worm early embryonic and adult gonad tissues to fly early embryonic and female adult stages, ii) alignment of fly ovary tissues to worm early embryonic and adult stages, iii) alignment of fly cell lines and gonad tissues to worm early embryos and gonad tissues, and iv) alignment of fly head and CNS tissues with worm neuron cells.

This study provides the first comprehensive transciptome-level comparison of multiple developmental stages, tissues and cells between *D. melanogaster* and *C. elegans*, and it revealed a few previously unknown connections (i.e., alignments) between developmental stages and tissues/cells both within and between the two species. The next step is to conduct a functional study to better understand the underlying molecular biology mechanisms that lead to these observed alignments. One big obstacle arises from inconsistencies in gene ontology (GO) [104] and the discrepancy between the GO vocabulary annotated for the two species [105], making the comparative functional study a difficult task. We have some preliminary results on aligning fly and worm developmental stages by using GO instead of orthology to link the two species, i.e., two stages whose associated genes have high dependence in their corresponding GO terms will be aligned. Figure 4.8 shows that the within-species alignments by GO terms are similar but much noisier versions of our results in Figure 3A and Figure 4.4A. However, the between-species alignment by GO terms (Figure 4.8C) only contains scarce signals, because of the dissimilarity of GO vocabulary for the two species. Bearing the above issues in mind, we provide a complete list of associated genes for fly and worm developmental stages, tissues and cells (Dataset S4.3 and Dataset S4.4) as a resource for future functional study. Since splicing regulation is key step in transcription and plays significant roles in an organisms development and cell/tissue differentiation [106, 107], it is also important to consider splicing regulation in refining the alignment results of different stages, tissues and cells.

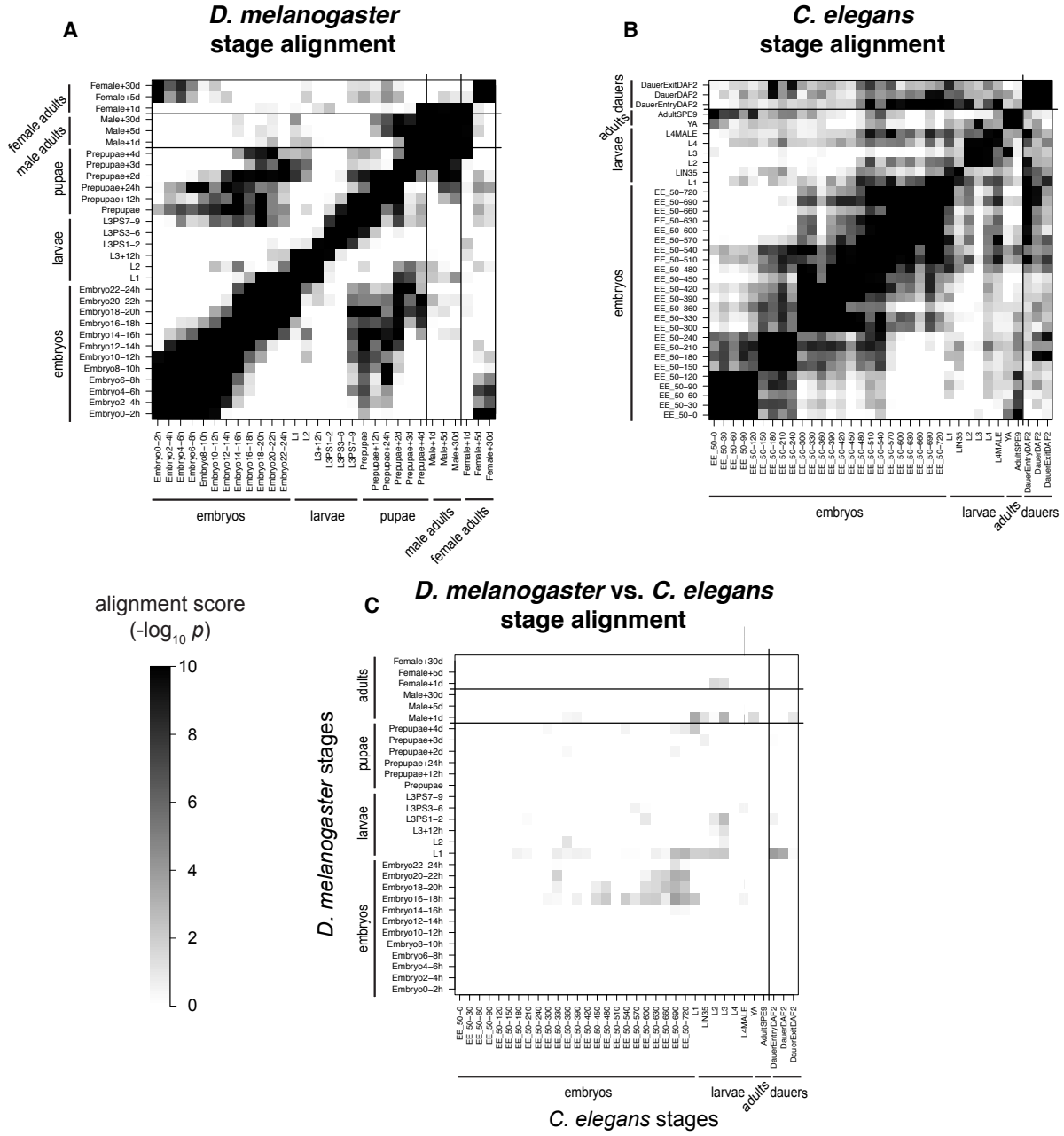

Figure 4.8: Preliminary stage alignment results based on Gene Ontology (GO) [104]. (A) Stage alignment within *D. melanogaster*. (B) Stage alignment within *C. elegans* stages. (C) Stage alignment between *D. melanogaster* and *C. elegans*. For every pair of stages (both within-species and between-species), a hypergeometric test was used to test the dependence of GO terms of their stage-associated genes. Leaf Biological Process (BP) GO terms were used in this analysis.

4.4 Materials and Methods

4.4.1 Estimating Gene Expression in Developmental Stages and Tissues/Cells

Cufflinks [16] (version 1.3.0, supplied with reference annotation, i.e. using "-G" option) was used to estimate the expression of 13,781 *D. melanogaster* protein-coding genes in 30 developmental stages, 29 tissues and 19 cell lines, and the expression 20,389 *C. elegans* protein-coding genes in 35 developmental stages and 18 tissues/cells. Gene annotations are from Ensembl assembly 66 [98] (i.e., BDGP 5.64 for *D. melanogaster* [97] and WS 220 for *C. elegans* [99]). All the gene expression estimates returned by Cufflinks are in FPKM (fragments per kilobase of transcript per million mapped reads) units. Hence, every fly and worm gene has one FPKM value per developmental stage/tissue/cell.

4.4.2 Identification of Stage/Tissue/Cell-Associated Genes

We use the identification of stage-associated genes for fly stages as an example. For every fly gene, suppose its expression estimates (in FPKM units) in the 30 developmental stages are $e_1, \ldots, e_{30}$. We normalize them as $z_1, \ldots, z_{30}$, where $z_i = \frac{e_i - \bar{e}}{s}$, $i = 1, \ldots, 30$, $\bar{e} = \frac{1}{30}\sum_{i=1}^{30} e_i$ and $s = \sqrt{\frac{1}{29}\sum_{i=1}^{30}(e_i - \bar{e})^2}$. Note that e_i represents the absolute expression level of the gene in stage i, and z_i represents the relative expression level of the gene in stage i as compared to other stages. For every fly stage, we would like to select the fly genes that have high relative expression and absolute expression distinguishable from background noise at that stage. The selection threshold is used in this manuscript is $z_i > 1.5$ and $e_i > 1$. If a gene satisfies this threshold, it will be selected as an associated gene of stage i. Based on our experience with gene expression estimates by Cufflinks, $e_i > 1$ is a reasonable cutoff to distinguish real gene expression signal from background noise. We tried two other thresholds on the relative gene expression: $z_i > 1.2$ and $z_i > 1.8$, and found the alignment results very robust to the three thresholds, suggesting that $z_i > 1.5$ is a reasonable threshold.

For worm developmental stages, we used the same method and threshold to select their stage-associated genes. For fly tissues/cell lines (and also worm tissues/cells), we treated them like developmental stages in selecting their tissue/cell-associated genes. Hence, for every fly/worm stage/tissue/cell, a subset of protein-coding genes that are highly expressed but not always highly expressed in other stages/tissues/cells are selected as its associated genes.

4.4.3 Hypergeometic Testing in Stage/Tissue/Cell-Alignment within a Species

Given two stages, or a stage and a tissue/cell, or two tissues/cells of the same species (i.e., *D. melanogaster* or *C. elegans*), we aligned them by testing the dependence of their associated

genes, say, gene sets A and B. We regarded all the protein-coding genes of the species as the population, and regarded associated-gene sets A and B as two samples drawn from the population. The null hypothesis to be tested against is that A and B are two independent samples from the population; the alternative hypothesis is that A and B are dependent samples. This becomes a standard hypergeometric test, and the test statistic is number of genes shared by A and B. The larger the test statistics is, the more likely the null hypothesis will be rejected. The p-value of the test statistic is calculated as

$$p = \sum_{i=|A\cap B|}^{\min(|A|,|B|)} \frac{\binom{n}{i}\binom{n-i}{|A|-i}\binom{n-|A|}{|B|-i}}{\binom{n}{|A|}\binom{n}{|B|}},$$

where n is the total number of protein-coding genes, and $|A|$, $|B|$ and $|A\cap B|$ are the number of genes in gene sets A, B and $A\cap B$. Hence, for any two stages, or a stage and a tissue/cell, or two tissues/cells, the p-value indicates the extent of their dependence, in other words, the strength of their alignment. Due to the multiple testing issue, we corrected the p-value by Bonferroni correction:

$$\text{Bonferroni corrected } p\text{-value} = p\text{-value} \times (\text{number of alignments}).$$

In the alignment of the 30 developmental stages within fly, the number of alignments is $30 \times 30 = 900$. We then defined the alignment score as

$$\text{alignment score} = -\log_{10}(\text{Bonferroni corrected } p\text{-value}),$$

and summarized the alignment scores of all pairwise alignments into a matrix. If rows or columns of the matrix correspond to developmental stages, they will be ordered by the temporal order; otherwise, if rows or columns correspond to tissues/cells, they will be grouped by hierarchical clustering on the matrix. The ordered matrix will then be presented by a heatmap (e.g. Figure 4.3A-D and Figure 4.4A-C) to illustrate alignment patterns.

4.4.4 Hypergeometic Testing in Stage/Tissue/Cell-Alignment between Two Species

Given two stages, or a stage and a tissue/cell, or two tissues/cells from two different species (i.e., *D. melanogaster* and *C. elegans*), we aligned them by testing the dependence of orthologs in their associated genes, say, fly gene set F and worm gene set W. We restricted both F and W to the associated genes that have orthologs in the other species. We regarded the 11,403 ortholog pairs between the two species (Table 4.2) as the population, represented by a two-column array of 11,403 rows:

ortholog pair type (fly-worm)		# fly genes	# worm genes	# pairs
1-1		3,131	3,131	3,131
1-many	1-2	310	620	620
	1-3	79	237	237
	1-4	37	148	148
	1-≥5	53	465	465
many-1	2-1	618	309	618
	3-1	234	78	234
	4-1	76	19	76
	≥5-1	262	32	262
many-many	2-2	132	132	264
	2-≥3	76	154	308
	≥3-2	136	60	272
	≥3-≥3	323	354	4,768
total		5,467	5,739	11,403

Table 4.2: Summary of *D. melanogaster* and *C. elegans* orthologs[a]

[a]from modENCODE prediction of fly-worm orthologs (`http://compbio.mit.edu/modencode/orthologs/modencode-orths-2012-01-30/ensembl-v65/modencode.merged.orth.txt.gz`)

$$\begin{array}{ccc} \text{fly gene} & & \text{worm gene} \\ f_1 & — & w_1 \\ \vdots & — & \vdots \\ f_{11,403} & — & w_{11,403} \end{array}$$

where f_i and w_i are the fly and worm genes in the ith ortholog pair. Please note that there exist repetitive genes in $\{f_1, \ldots, f_{11,403}\}$ and $\{w_1, \ldots, w_{11,403}\}$ due to the existence of 1-to-many, many-to-1 and many-to-many ortholog pairs. Since F and W contain no repetitive genes, we define $F' = \{f_i : f_i \in F, i = 1, \ldots, 11,403\}$ and $W' = \{w_i : w_i \in W, i = 1, \ldots, 11,403\}$ as alternative versions of F and W with repetitive genes. We then regarded F' as a sample from $\{f_1, \ldots, f_{11,403}\}$ (i.e. the fly gene part of the population) and W' as a sample from $\{w_1, \ldots, w_{11,403}\}$ (i.e. the worm gene part of the population). Because of the one-to-one correspondence between $\{f_1, \ldots, f_{11,403}\}$ and $\{w_1, \ldots, w_{11,403}\}$, we can consider F' and W' as two samples from the same population.

The null hypothesis to be tested against is that F' and W' are independent samples from the population; the alternative hypothesis is that F' and W' are dependent samples. This becomes a hypergeometric test setting, and the test statistic is the number of ortholog pairs

existing between F' and W', defined as T. The larger the test statistics is, the more likely the null hypothesis will be rejected. The p-value of the test statistic is calculated as

$$p = \sum_{i=T}^{\min(|F'|,|W'|)} \frac{\binom{11,403}{i}\binom{11,403-i}{|F'|-i}\binom{11,403-|F'|}{|W'|-i}}{\binom{11,403}{|F'|}\binom{11,403}{|W'|}},$$

where $|F'|$ and $|W'|$ are the number of genes (including repetitive ones) in gene sets F' and W'. Hence, for any two stages, or a stage and a tissue/cell, or two tissues/cells from two different species, the p-value indicates the extent of their dependence, in other words, the strength of their alignment. Then similar to the alignment within a species, we addressed the multiple testing issue by correcting the p-values by Bonferroni correction and subsequently calculated alignment scores as $-\log_{10}$(Bonferroni corrected p-value). The alignment result is also summarized in a matrix, where hierarchical clustering is applied to order the tissues/cells as rows or columns, and finally represented by a heatmap (Figure 4.5A-D) [108, 109].

4.5 Acknowledgements

This work was supported by NIH/NHGRI U01 HG004271 to Dr. Susan Celniker and NIH/NHGRI RC2 HG005639 to Dr. Manolis Kellis. I would like to thank Dr. Haiyan Huang, Dr. Peter J. Bickel and Dr. Steven E. Brenner as the co-authors of this work. We thank Dr. Susan E. Celniker, Dr. Robert Waterston, Dr. Mark B. Gerstein, Dr. Roger Hoskins and Dr. LaDeana Hillier for their insightful comments. We would also like to thank the modENCODE consortium for their data and support.

Chapter 5

Conclusions

5.1 Summary

In this thesis, we investigated three important biological questions by applying statistical and computational approaches to analyzing high-throughput genomic data.

In Chapter 2, we developed a statistical software package "SLIDE" (Sparse Linear modeling of RNA-Seq data for Isoform Discovery and abundance Estimation) for discovering and quantifying mRNA isoforms from next-generation RNA-Seq data. Unlike other isoform discovery methods (e.g. Cufflinks [16] and Scripture [30]) that use deterministic graphical models, we considered the stochastic nature of RNA-Seq data and formulated the isoform discovery and abundance estimation problem using a linear regression framework, where observations are RNA-Seq read counts (numbers of short sequences generated from mRNA molecules) and parameters are unknown isoform abundances. The design matrix models stochastic relationships between reads and isoforms: probabilities of reads coming from different possible isoforms. For statistical inference, SLIDE takes a two-step approach: it first discovers isoforms from sparse estimates of the linear model by L_1 regularized regression; it then estimates the abundance of discovered isoforms from the reduced linear model by non-negative least squares. This linear regression framework is flexible to account for biases in RNA-Seq data (e.g. GC-content bias [17, 19]) and to incorporate other transcriptomic data resources (e.g. EST [33], CAGE [32], and RACE [31]). By simulation and real data analysis, we demonstrated SLIDE as a useful tool for discovering and quantifying mRNA isoforms from RNA-Seq data and showed its better performance compared to major competitors in the isoform discovery field.

In Chapter 3, we corrected system-wide protein abundance estimates by using previously reported individual protein abundance measurements,and found that the median protein abundance in mammalian tissue culture cells increased from 8,000 - 16,000 molecules per cell (estimated by system-wide label free quantification) to 170,000 molecules per cell, a more reasonable number expected from literatures. We then used the corrected protein abundance estimates to re-determine the contribution of transcription to the variance of

protein abundances. System-wide surveys suggest that differences in mRNA expression between genes explain only 10-40% of the differences in protein levels. We found, however, that mRNA levels explained a higher percentage of the variance in protein levels, by using our corrected protein abundance estimate and taking direct measurements of experimental error into account. We estimated that mRNA levels explain at least 56% of the differences in protein abundance for the 4,212 genes detected by Schwanhausser et al [11]. By in addition modeling all genes' expression, we show that under reasonable assumptions mRNA levels can explain at least 65% of protein levels for genes that are expressed and 100% for genes that are not expressed. Separately, we employ a second strategy to determine the contribution of mRNA levels to protein expression. This shows that the variance in translation rates directly measured by Ingolia et al. [43] is 4.6 fold less than the variance inferred by Schwanhausser et al. and that based on this mRNA levels are expected to explain $\sim 75\%$ of the variance in protein levels for the 4,212 detected genes and $\sim 82\%$ for all expressed genes. While the magnitude of our differently derived estimates vary, all suggest that the previous studies have significantly underestimated the importance of transcription.

In Chapter 4, we undertook a comparison of the developmental stages, tissues and cells of two model organisms, seeking commonalities in orthologous genes transcription. Our approach centers on using orthologous genes to link the two organisms, and finding stage/tissue/cell-associated genes to represent transcription in every developmental stage, tissue and cell. For every stage/tissue/cell in each organism, stage/tissue/cell-associated genes are selected as those highly expressed in that sample (i.e., a stage, or a tissue, or a cell) but not always highly expressed in all the other samples. We tested the dependence of a pair of fly and worm samples (in terms of orthologous gene expression) by using an overlap statistic, which is the number of orthologous gene pairs in their associated genes. Samples that exhibit statistically significant dependence are called "aligned". This alignment strategy was first applied to pairwise stages/tissue/cells within fly and worm respectively, and the within-species alignment results are expected and reasonable based on previous reports and biological knowledge, thus supporting the validity of our approach. Our next alignment results of stages/tissue/cells between fly and worm are the first findings regarding a comprehensive comparison of multiple biological samples in the two model organisms. Our most important discovery is the two parallel alignment patterns between fly and worm developmental timecourses. This result implies that there may exist conservation in post-embryonic development even between evolutionarily-distant species. We suggest, in addition, that in comparing related biological samples, using subsets of genes that capture transcriptional characteristics of the samples is a more effective approach than widely-used correlation-based approaches, because the genes with constantly high or low expression levels in all the samples contain no useful information for the comparison.

5.2 Future Directions

In this section, two future directions will be proposed as possible extensions of the works described in this thesis, with the goal to further address important problems in the interdisciplinary field of statistics and genomics.

5.2.1 Joint Modeling of Multiple RNA-Seq Samples for Isoform Discovery and Abundance Estimation

Given the availability of multiple biological/technical replicate RNA-Seq data produced by the ENCODE [83] and modENCODE [13] consortia, it becomes necessary to extend our SLIDE method (Chapter 2) to a newer version that can achieve better accuracy by using replicate data.

SLIDE was originally designed to discover and quantify mRNA isoforms from a single RNA-Seq dataset. In the case where multiple biological or technical replicate RNA-Seq data from the same sample exist, the simplest and most common way of handling the replicates is to pool their reads together into one dataset and subsequently input the so-called "pooled data" into any downstream algorithms. Another common way of using replicates is to process each replicate separately with a chosen algorithm and then combine the outputs of different replicates into one final output. Either of the above two ways has its advantages and drawbacks. The former way, say "pooling method", combines RNA-Seq reads from different replicates into a pooled dataset, which is expected to have less non-systematic noise and bias than an individual replicate has, if all replicates are correctly produced under similar experimental conditions. However, if one or more replicates are much noisier than the rest of replicates, or if a small number of replicates were produced under a different condition that experimenters were unaware of, this pooled dataset may be biased by those "outlier" replicates, whose noise may mask useful information in the majority of replicates with biological interest. The latter way, say "individual method", treats replicates as independent datasets, and thus "outlier" replicates would not interfere with the majority of replicates. However, individual outputs may be biased by their own non-systematic noise, and thus it is not an easy task to combine those outputs into a more biologically meaningful output.

After realizing the disadvantages of the above two common methods, we are motivated to find a better way of using replicates in the SLIDE method. Here we propose a possible way of extending SLIDE for multiple replicate RNA-Seq data.

We increase the dimensionality of observations $(b_j)_{j=1}^J$ and parameters $(p_k)_{k=1}^K$ in Equation (2.1), i.e.,

$$\left[b_{1j}, \cdots, b_{(m+1)j}\right] = \left[\sum_{k=1}^{K} F_{jk}p_{1k}, \cdots, \sum_{k=1}^{K} F_{jk}p_{(m+1)k}\right] + \left[\epsilon_{1j}, \cdots, \epsilon_{(m+1)j}\right] \quad j = 1, \cdots, J, \tag{5.1}$$

where m is the total number of replicates, b_{ij} is the observed proportion of reads in the jth bin in the ith replicate, F_{jk} is the conditional probability same as in Equation (2.1), p_{ik} is

the proportion of the kth isoform in the ith replicate, and ϵ_{ij} is the error term with mean 0. The $(m+1)$th replicate is the "pooled data". A simple and common way of pooling multiple RNA-Seq replicates into one dataset is to calculate the *pooled bin proportions* as

$$b_{(m+1)j} = \frac{1}{m}\sum_{i=1}^{m} b_{ij}, \quad j = 1, \ldots, J. \tag{5.2}$$

In the hope of using individual replicates to help reduce possible biases in the pooled data, we can jointly estimate the isoform proportions in Equation (5.1) as

$$[\hat{\mathbf{p}}_1, \cdots, \hat{\mathbf{p}}_{m+1}] = \arg\min_{\mathbf{p}_1,\cdots,\mathbf{p}_{m+1}} \left(\sum_{i=1}^{m+1}\sum_{j=1}^{J}(b_{ij} - \mathbf{F}_j\mathbf{p}_i)^2 + \lambda_1 \sum_{i=1}^{m+1}\sum_{k=1}^{K}\frac{|p_{ik}|}{n_k} \right.$$
$$\left. +\lambda_2 \sum_{i=1}^{m}\sum_{k=1}^{K}|p_{(m+1)k} - p_{ik}| \right), \ \ s.t.\ p_{ik} \geq 0, \tag{5.3}$$

where $\hat{\mathbf{p}}_1, \cdots, \hat{\mathbf{p}}_m$ are isoform proportion estimates from Replicate $1, \cdots, m$, and $\hat{\mathbf{p}}_{m+1}$ are isoform proportion estimates from the pooled data. The estimate of our interest is $\hat{\mathbf{p}}_{m+1}$. By adding the second penalty term to the objective function, we intend to have the estimates from the pooled data similar to the majority of replicates, and dissimilar to the outlier replicates. That is, we expect the estimates $\hat{\mathbf{p}}_{m+1}$ from Equation (5.3) to be more robust to "outlier" replicates and thus more accurate than estimates from the "pooling method" and the "individual method".

5.2.2 Further Studies on modENCODE Timecourse Data

In our current work on the comparison of *D. melanogaster* and *C. elegans* developmental stages (Chapter 4), we did not use the temporal information of stages. We would like to further develop statistical methods for timecourse alignment, by regarding gene expression profiles as functional data with different time scales. The goal is to find a partial alignment between timecourses of different species, and to identify important stages that contribute most to the alignment. This study will have a broad interest beyond the comparison of two model organisms, as it may be applied to aligning two relevant groups of functional data with different x-axis scales.

Another question of biological interest is to incorporate alternative splicing information into aligning different *D. melanogaster* and *C. elegans* stage/tissue/cell samples. It will be interesting and important to study whether alternative splicing patterns in the two species have similar dynamics throughout their development, and whether our observed alignment patterns in Chapter 4 remain the same after considering alternative splicing. New statistical methods are needed to address such issues.

5.3 Discussion

With the rapid development of high-throughput genomic technologies, many previously unsolved or controversial biological questions can now be answered or addressed from a new perspective. To answer or address these questions, statistical methods and analysis are necessary tools. In this thesis, we described three interesting examples that demonstrate the power of combining statistics and high-throughput genomic data to study important questions in the genomics field. Given the enormous amount of high-throughput genomic data and a large number of existing and new biological questions, we believe that applying statistics to analyzing high-throughput genomic data is an attractive topic and will become an increasingly prominent direction in interdisciplinary research.

Appendix A

Supplementary Material for "Sparse Linear Modeling of Next-Generation mRNA Sequencing (RNA-Seq) Data for Isoform Discovery and Abundance Estimation"

A.1 Linear Modeling of RNA-Seq data

Linear modeling of paired-end RNA-Seq data has been discussed in Section 2.4. The main points include (i) the definition of *paired-end bins* to summarize the key information in RNA-Seq data for isoform discovery, (ii) the enumeration of all possible isoforms from defined *subexons*, (iii) the modeling of conditional probabilities of observing reads in different bins given an isoform, and (iv) the construction of a linear model to estimate isoform proportions from observed bin counts.

Below we present more details about modeling the fragment length distribution in **F** and construction of linear model for single-end data.

A.1.1 The Fragment Length Distribution

Modeling the cDNA fragment length distribution is a key part in constructing the design matrix **F**F of the linear model. Truncated Exponential is a reasonable candidate for the distribution, based on a Poisson point process assumption on a fragment's 3' end with the 5' end fixed and a size selection step in RNA-Seq protocols. Another widely used candidate in existing RNA-Seq tools is Normal distribution [16]. To evaluate the two distributions, we compared them with empirical distributions of cDNA fragment lengths in paired-end RNA-Seq data. However, actual fragment lengths are unknown in genes exhibiting alternative

splicing events, thus posing difficulties in obtaining the empirical distributions. To tackle this problem, a conservative solution is to calculate an empirical length distribution of cDNA fragments with both ends in the same *subexon* where no alternative splicing occurs. The good side of this solution is that the fragment lengths used in the calculation are actual, but the downside is that some long fragments across exons are not considered. Another solution is to calculate an empirical length distribution based on cDNA fragments in genes with no alternative splicing events in the UCSC *Drosophila melanogaster* (September 2010) annotation [29]. This solution has the advantage of observing all sorts of fragment lengths, but its disadvantage is that wrong fragment lengths may be used if the annotation is incomplete. We employed both solutions to calculate the empirical distributions from dataset 1 (Table 2.1), and plotted them against either truncated Exponential or Normal distribution in Q-Q plots (Figure A.1). Parameters in the truncated Exponential and Normal distributions are chosen in such a way that both distributions have the same mean and variance as in the empirical distribution. Q-Q plots in Figure A.1 show that both truncated Exponential and Normal distributions are reasonable approximations of the fragment length distribution.

A.1.2 Linear Modeling of Single-End RNA-Seq Data

For single-end RNA-Seq data, we can derive a similar linear model to the one used for paired-end data (Equation 2.3). First, we enumerate possible isoforms in the same way as for the paired-end data, and categorize reads into *single-end bins*, defined as two-dimensional vectors indicating subexon indices of the reads. For example, single-end bin (i, j) contains reads whose 5' and 3' ends are in subexon i and j, respectively. A *single-end bin count* is defined as the number of reads in that bin. Bin counts of every gene are normalized as *bin proportions*, denoted by $\mathbf{b}$. Second, we construct a linear model to estimate isoform proportions $\mathbf{p}$ from observed single-end bin proportions, with a design matrix $\mathbf{F}$ as the conditional probabilities of observing reads in different single-end bins given an isoform. The modeling and calculation of the conditional probabilities for single-end data are similar to those for paired-end data in the main paper. We consider a single-end bin as equivalent to a combination of multiple paired-end bins. For example, in a two-subexon gene, reads in single-end bin (1,1) correspond to paired-end reads in bins (1,1,1,1), (1,1,1,2), and (1,1,2,2). So the conditional probability of observing reads in single-end bin (1,1) given an isoform equals to the sum of conditional probabilities of observing reads in each of the three paired-end bins given the same isoform. In general, we calculate the conditional probability of observing reads in single-end bin j given isoform k as $\sum_{r \in S_j} F'_{rk}$, where S_j is the set of paired-end bins corresponding to the single-end bin j, and F'_{rk} is the conditional probability for paired-end data whose calculation has been described in details in the main paper. Last, we write a linear model in the same formula as in Equation 2.3.

For combined paired-end and single-end data, we can simply construct a linear model by catenating the observation vectors and combining the design matrices by rows in the linear models for paired-end and single-end data, respectively. Hence, the linear model used in SLIDE (Sparse Linear modeling of RNA-Seq data for Isoform Discovery and abundance

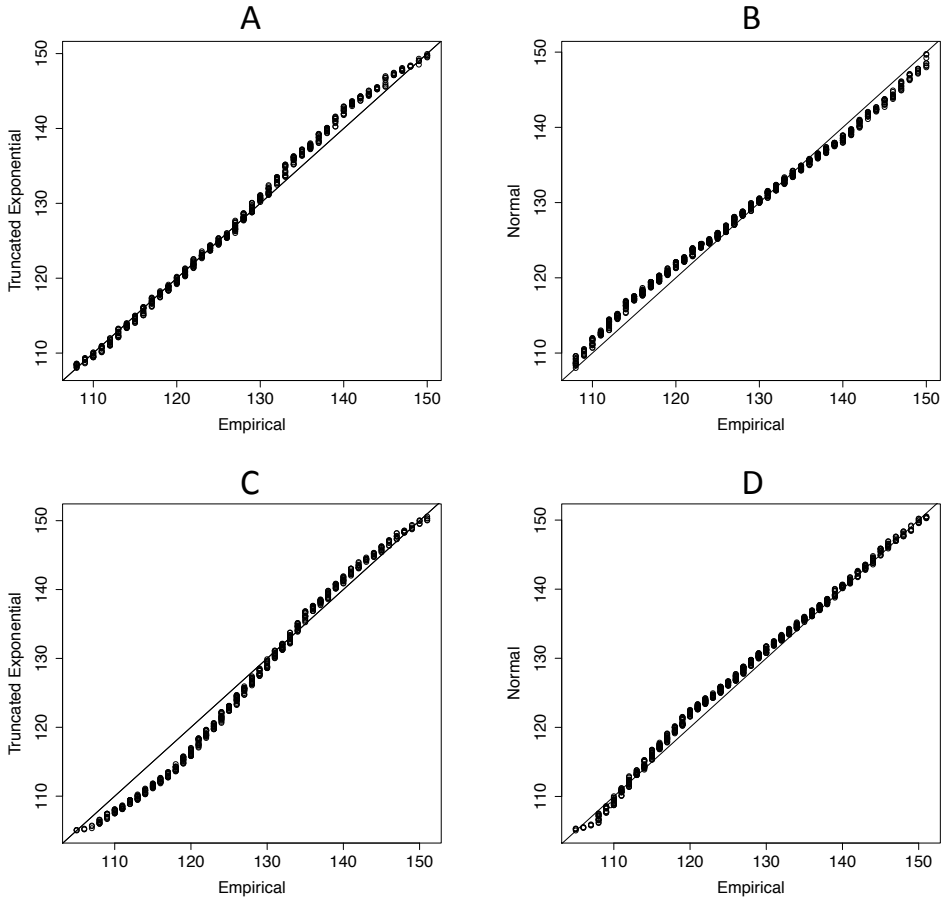

Figure A.1: Q-Q plots of modeled vs. empirical fragment length distribution on dataset 1 (Table 2.1). Note that only the fragment lengths between the 5% and 95% percentiles of the empirical distribution are used to construct the Q-Q plots, because extremely long or short fragments may be results of mapping errors. (A) Q-Q plot of truncated Exponential distribution vs. empirical length distribution of cDNA fragments within single-exon genes. (B) Q-Q plot of Normal distribution vs. empirical length distribution of cDNA fragments within single-exon genes. (C) Q-Q plot of truncated Exponential distribution vs. empirical length distribution of cDNA fragments within single-isoform genes. (D) Q-Q plot of Normal distribution vs. empirical length distribution of cDNA fragments within single-isoform genes.

No. of subexons	Total no. of genes	No. of genes with unidentifiability issue before the preselection procedures	No. of genes with unidentifiability issue after the preselection procedures
3	295	204 (69.2%)	1 (0.3%)
4	237	228 (96.2%)	198 (83.5%)
5	165	165 (100%)	155 (93.9%)
6	142	142 (100%)	137 (96.5%)
7	82	82(100%)	80 (97.6%)
8	72	72 (100%)	70 (97.2%)
9	56	56 (100%)	55 (98.2%)
10	35	35 (100%)	35 (100%)

Table A.1: Number of genes with unidentifiability issues before and after preselection procedures

Estimation) can accommodate for different types of RNA-Seq data: paired-end, single-end, or both.

A.1.3 Identifiability and Pre-Selection Procedures

To avoid the unidentifiability issue due to collinearity in the linear model (Equation 2.2), we applied a preselection procedure: Only isoforms whose all subexon junctions have been observed are selected as candidates; for genes with more than two subexons, single-subexon isoforms are excluded from the candidates because of their rare existence. With this procedure, the number of parameters for an n-subexon gene can be reduced from $2^n - 1$ to a significantly smaller number.

About the observations, there are frequently false zero counts of *junction-end bins*. We define *junction-end bins* as bins that include paired-end reads with at least one end across exon junctions [e.g., junction-end bins (1,1,1,2) and (1,2,2,2) include paired-end reads with one end covering the junction between subexons 1 and 2, whereas bin (1,1,2,2) is not a junction-end bin]. When bin (1,1,2,2) has positive counts, the expected counts of bins (1,1,1,2) and (1,2,2,2) should be positive, too; however, due to the difficulty of mapping junction reads, junction-end bins (1,1,1,2) and (1,2,2,2) are often observed with false zero counts. Thus, we exclude false zero junction-end bin proportions from the observations.

As an illustration of the effects of such preselection procedures, we calculate the numbers of genes with unidentifiability issues in their linear models (i.e., $\text{rank}(\mathbf{F}) < K$ in Equation 2.1) before and after the preselection procedures for every group of n-subexon genes ($n = 3, \ldots, 10$). The numbers are summarized in Table A.1, which shows that the preselection procedures have effectively overcome the unidentifiability issue for genes with three subexons and alleviated the problem for a few genes with more subexons. However, the percentage of genes with unidentifiablity issues remains high after the preselection procedure for genes with more than three subexons; we see that the sparse estimation in SLIDE is still necessary.

A.1.4 L_1 vs. L_0 Regularization

In sparse estimation, L_1 penalty in Lasso is linear and ensures convexity of the objective function (Equation 2.4). It also has the convexity property in logistic and Poisson regressions. Lasso does variable selection and shrinkage, thus permitting isoform discovery in SLIDE. L_0 penalty is also a possible choice for sparse estimation. It was reported that L_0 penalty can lead to a sparser model when the number of variables (e.g., the number of isoform candidates) is far larger than the number of relevant variables (e.g., the number of existing isoforms), whereas L_1 penalty in Lasso has to use a large λ to screen out spurious variables and causes biases in retained variables [34, 110]. However, L_0 regularization is computationally disadvantageous because it makes the optimization problem nonconvex, and it has been shown that L_1 is a good surrogate for L_0 in many cases. In computational biology, L_1 regularization is shown to be a good approach for high-dimensional and potentially sparse data [38]. In our case, SLIDE does isoform discovery and abundance estimation in two steps, so the biased estimates of isoforms in the discovery step would not affect the subsequent abundance estimation step as long as true isoform estimates are not shrunk to zeros by Lasso. This is different from IsoLasso and NSMAP, which use one-step sparse estimation for simultaneous isoform discovery and abundance estimation [27, 36]. Moreover, in our estimation, the existence of n_k (the number of exons in the kth isoform) in the penalty term would reduce the difference between L_1 and L_0 regularization. Therefore, L_1 regularization is a reasonable choice for our sparse estimation.

A.1.5 Selection of the Regularization Parameter in Sparse Estimation

The selection of the regularization parameter λ (Equation 2.4) is by a stability criterion that aims to return the most stable results over different runs of estimation [37]. Because genes of the same number of subexons have similar $\dim(\mathbf{p})$ and $\dim(\mathbf{b})$ in Equation 2.4 of the main paper, we decided to group genes by their numbers of subexons n and select an optimal $\lambda^{(n)}$ for each group from 16 candidate values $(\lambda_i)_{i=1}^{16}$ (see Table 2.3). This grouping is particularly advantageous for selecting λ for lowly expressed genes, whose signal-to-noise ratios are low. Highly expressed gene signals can counteract the noise in the lowly expressed genes of the same group.

Suppose that there are $m^{(n)}$ genes with n subexons, $n = 3, \cdots, 10$. The selection procedures following the stability criterion are as follows.

1. For the rth gene with n subexons, $r = 1, \cdots, m^{(n)}$, use $\lambda = \lambda_i$, $i = 1, \cdots, 16$ in Equation 2.4 to estimate $\hat{\mathbf{p}}$ for 50 runs. In each run, use randomly sampled one half of the reads in the gene as input into SLIDE. Define q_{irk} as the proportion of runs in which $\hat{p}_k > 0$. Define $\bar{q}_{ir} = \frac{\sum_{k=1}^{K} q_{irk}}{\sum_{k=1}^{K} I(\hat{p}_k > 0 \text{ in some runs})}$.

2. Calculate the average of $\bar{q}_{ir}$ over the $m^{(n)}$ genes as $\tilde{q}_i = \frac{1}{m^{(n)}} \sum_{r=1}^{m^{(n)}} \bar{q}_{ir}$.

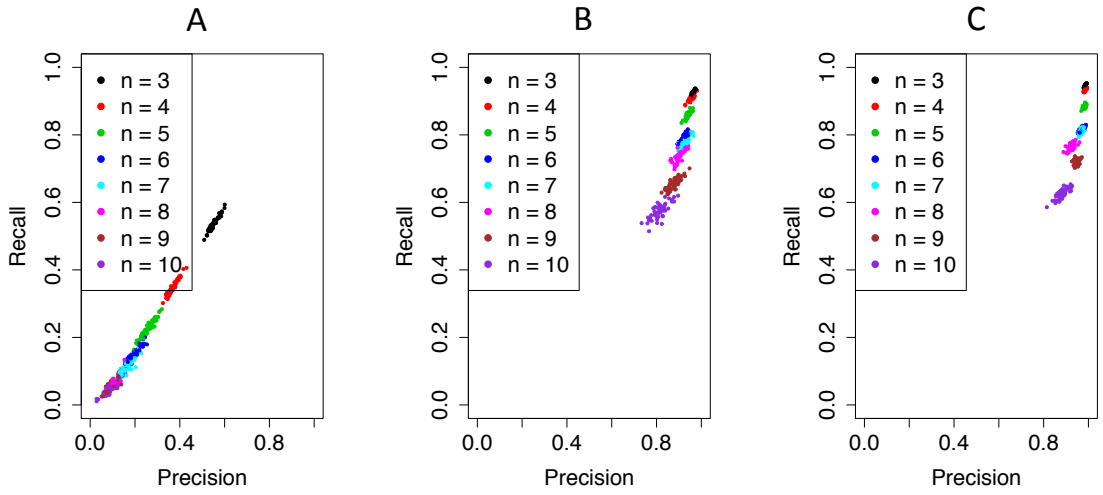

Figure A.2: Precision and recall rates of SLIDE on simulated data with different read coverages. (A) Read coverage is 10 reads per kilobase. (B) Read coverage is 50 reads per kilobase. (C) Read coverage is 100 reads per kilobase.

3. Choose $\lambda^{(n)}$ as λ_{i^*}, where $i^* = \arg\max_i \tilde{q}_i$.

The selected $\lambda^{(n)}$ for datasets 1-4 (Table 2.1) and the simulation data are in Table 2.3.

A.2 More Simulation Studies

A.2.1 Simulation Studies with Different Read Coverages

To study the isoform discovery accuracy of SLIDE in lowly expressed genes, we did a simulation study with three different read coverages: (i) 10 reads per kilobase of an annotated gene, (ii) 50 reads per kilobase of an annotated gene, and (iii) 100 reads per kilobase of an annotated gene. The simulated reads are paired-end with 37-bp length in each end. Precision and recall rates of SLIDE using the simulated data are summarized in Figure A.2, which shows that SLIDE has improved isoform discovery accuracy as the read coverage increases, as we expected. The improvement is significant when the read coverage increases from 10 reads per kilobase to 50 reads per kilobase, and the improvement becomes less significant when the read coverage increases further to 100 reads per kilobase. Given that many paired-end RNA-Seq data have more than 10 million reads, 10 reads per kilobase would correspond to less than 1 RPKM (number of reads per kilobase per million of mapped reads) in those data. We note that a gene with such low read coverage and multiple exons is not likely to have all its exon junctions covered by reads, thus posing great difficulties on isoform discovery. As illustrated by this simulation study, SLIDE is robust to changes in gene expression levels when read coverage is beyond a certain threshold, and SLIDE has higher precision and recall

rates and lower estimation variance as read coverage increases. When gene expression is too low (e.g., 10 reads per kilobase), some exons or exon junctions would not be observed and the dimensionality of observations in the core linear model would be reduced, thus resulting in incorrect estimation results by SLIDE. At the read coverage of 10 reads per kilobase, we have tried other likelihoods (multinomial and Poisson) to model the responses (i.e., bin counts) in the linear model of SLIDE, but the precision and recall rates are similarly low (see Subsection A.2.2). [Please note that our Poisson regression has a similar objective function as the maximum-likelihood approach used in NSMAP [36] has in the optimization, except for differences in the design matrix and penalty term.] This missing data problem associated with lowly expressed genes is not unique to SLIDE, because to accurately recover missing reads from observed data remains a big challenge for current RNA-Seq isoform discovery and quantification methods. Because of data noise and biases introduced at many experimental steps of the current RNA-Seq protocol, it would be difficult to recover missing exons or junctions by statistical models.

A.2.2 Simulation Studies with Different Likelihoods in the Core Linear Model

To explore the effects of using different likelihoods in the generalized linear model of SLIDE (Equation 2.3), we tried three different likelihoods: Normal (the default), multinomial (logistic regression) and Poisson in the sparse estimation with simulated data. Reads were simulated under two read coverages, 10 and 100 reads per kilobase. Simulation settings are the same as described in the main paper. The results in Figure A.3 illustrate that in general, the three different likelihoods do not give very different results in both read coverages. Looking more closely, we find that using Normal likelihood at read coverage 10 reads per kb gives slightly higher precision and recall rates for genes with 3-4 subexons, and using Logistic regression at read coverage 100 reads per kb gives lower precision rates for genes with 3-5 exons. In our SLIDE model, it is naturally to assume that the expected bin counts are linear in isoform quantities and to use an identity link function (Normal likelihood). These exploration results confirm that Normal likelihood is a reasonable choice.

A.2.3 Effects of Isoform Similarity and Missing Annotations on Isoform Discovery

Similarity between different isoforms of the same gene would pose difficulties on isoform discovery. There are some cases where the isoform deconvolution is not identifiable because of the similarity between true isoforms [111, 112]. For example, when some isoforms are fragments of others in the true isoform set, there would usually be more than one possible set of isoforms that can explain the observed exon expression levels and exon junctions.

In situations that annotations have missing but truly expressed isoforms, there are two different cases. First, when missing isoforms have exons not included in annotated isoforms,

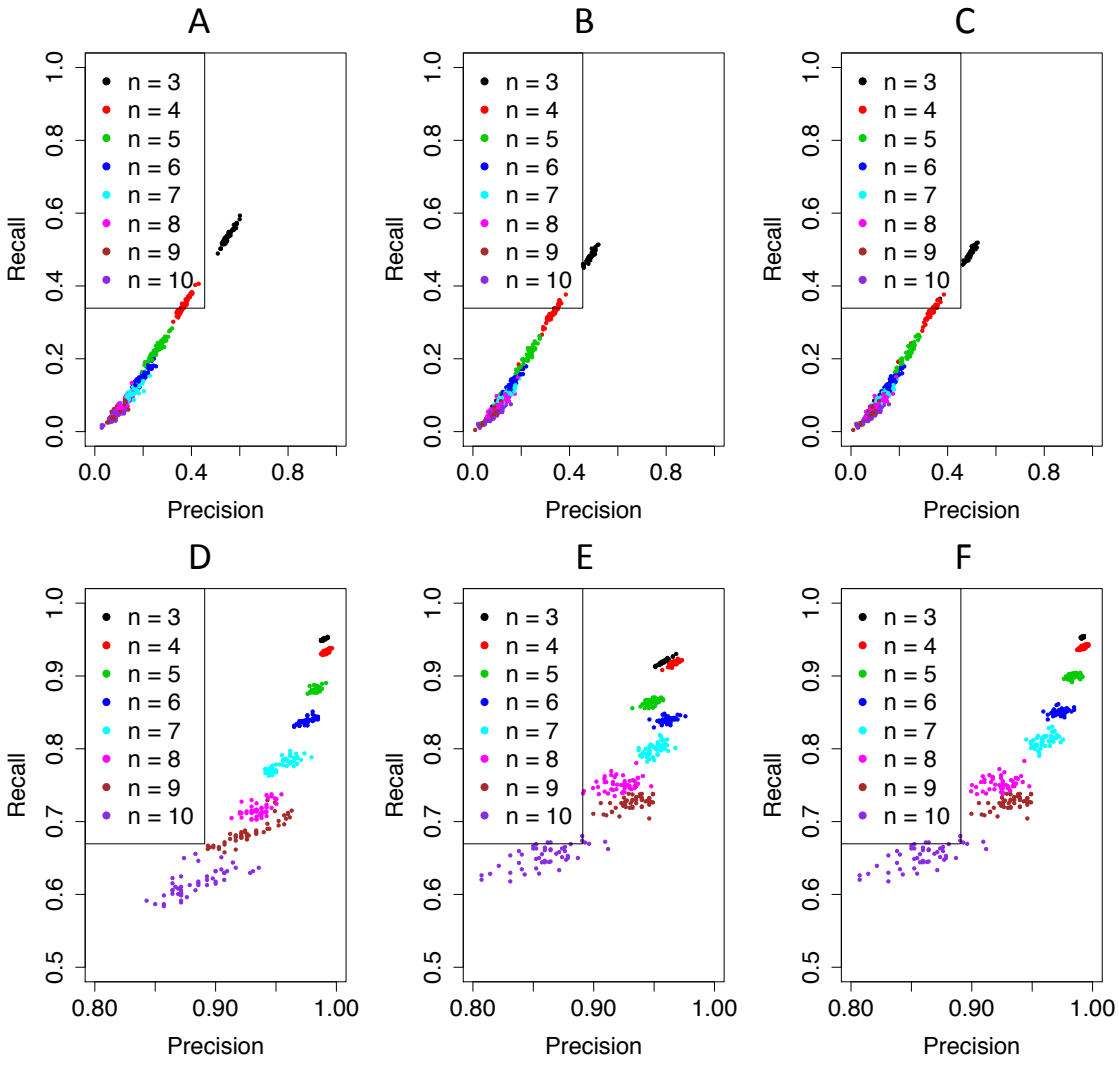

Figure A.3: Precision and recall rates of SLIDE using different likelihoods in simulation with two different read coverages. (A) Normal likelihood, (B) Poisson likelihood, and (C) multinomial likelihood (logistic regression) with read coverage 10 reads per kb. (D) Normal likelihood, (E) Poisson likelihood, and (F) multinomial likelihood (logistic regression) with read coverage 100 reads per kb.

although SLIDE is not designed to recover missing exons from data, it can solve this issue by using de novo exons assembled by other softwares [e.g., Cufflinks [16], Scripture [31]]. Second, when all the exons in missing isoforms are included in annotated isoforms, SLIDE can discover the missing isoforms with high accuracy, especially if every missing isoform has more than one unique splice junctions. In the difficult case where some missing isoforms are fragments of annotated isoforms and the isoform deconvolution is not identifiable, SLIDE would discover a set of longest isoforms with the highest probability among all the possible sets of isoforms. For example, we suppose that a three-exon gene has exon RPKMs 10, 20, and 10, respectively, and junction reads are observed between exons 1-2, and exons 2-3. In terms of the isoform deconvolution, there would be two possible sets of isoforms: (i) isoform (1,2,3) with RPKM 10 and isoform (2) with RPKM 10; or (ii) isoform (1,2) with RPKM 10 and isoform (2,3) with RPKM 10. In this case, SLIDE would favor the latter (set ii), which has a smaller penalty term. We design SLIDE to favor longer isoforms in the sparse estimation, by weighting each isoform abundance estimate with the inverse of its number of exons. This is based on our observations that most annotated isoforms contain many instead of few exons. In real data study, there are commonly observed 5' and 3' end biases in RNA-Seq data, that is, in our example above, even if the true isoform is (1,2,3), RNA-Seq read coverage in exons 1 and 3 is very likely to be lower than the read coverage in exon 2. To counteract the end biases in real RNA-Seq data, we allow SLIDE to favor isoforms with more exons in the sparse estimation. Therefore, SLIDE would find the longest isoform containing all the three exons unless the read coverage difference between exons 1 and 3 and exon 2 is significantly high.

We did simulation studies in the following three cases to illustrate the performance of SLIDE when annotations have missing isoforms but contain all the exons. In gene *RhoL*, there are four exons with lengths 379, 172, 286, and 204, respectively. The only annotated isoform is (1,2,3,4) that contains all four exons. In each of the following cases, we did 50 simulation runs with 500 paired-end 37-bp reads simulated in each run.

Case 1. Suppose that isoform (1,3,4) is missing in the annotation and its expression level is the same as that of isoform (1,2,3,4). We note that (1,3,4) contains a novel junction between exons 1 and 3 that is not in the annotated isoform (1,2,3,4). For all 50 runs, SLIDE correctly discovered both isoforms.

Case 2. Suppose that isoform (2,3,4) is missing in the annotation and its expression level is the same as that of isoform (1,2,3,4). We note that (2,3,4) is a fragment of the annotated isoform (1,2,3,4). For 49 out of the 50 runs, SLIDE correctly discovered both isoforms.

Case 3. Suppose that both isoforms (1,3,4) and (2,3,4) are missing in the annotation and both of their expression levels are the same as that of isoform (1,2,3,4). SLIDE correctly discovered all three isoforms in 18 runs. It missed isoform (2,3,4) in 18 runs, missed isoform (1,3,4) in 8 runs, and missed both in 6 runs.

From the results, we can see that it is more difficult to discover missing isoforms that are fragments of annotated isoforms, because SLIDE has to tackle end biases in real data. Nevertheless, these simulation results show that SLIDE has satisfactory performance in cases where annotations have missing isoforms.

A.3 More about mRNA Isoform Discovery on modENCODE Data

A.3.1 More about the Comparison between SLIDE and Cufflinks

In the main paper, we carried out three comparisons between SLIDE and Cufflinks from different perspectives. First, we compare the two methods in their default settings, where SLIDE uses genes and exons from UCSC annotations and Cufflinks uses its de novo assembled genes and exons (Figure 2.2B). In the evaluation step, we compare discovered isoforms by each method with isoforms in UCSC annotations. We call a discovered isoform and an annotated isoform matched if they have the same number of exons and all of their exons overlap. Thus, our evaluation scheme is not sensitive to exon boundaries as long as de novo assembled exons are in the same loci as annotated exons. We agree that this comparison is not fair for Cufflinks, but the results still reveal two main problems of the Cufflinks results: (i) Cufflinks splits a gene into multiple parts when few junction reads are observed between certain exons; (ii) Cufflinks merges two genes on opposite strands if they overlap because the read strand information is not properly considered. SLIDE does not have those two problems because it uses annotated gene boundaries that are mostly accurate. In the second comparison, we applied both methods to de novo assembled genes and exons by Cufflinks (i.e., to compare the isoform assembly performance of SLIDE and Cufflinks given the same set of genes and exons) (Figure 2.2C). The comparison results show that SLIDE and Cufflinks have similar precision and recall rates, which are, however, much lower than the precision and recall rates SLIDE had when using annotated genes and exons. We were concerned that the precision and recall rates in the second comparison might have been dominated by the de novo gene boundaries and exon loci that are different from the annotation. Therefore, we performed a third comparison between SLIDE and Cufflinks with only the genes whose de novo assembled exons agree with annotated exons in their loci. We found that the precision and recall rates of SLIDE are higher than those of Cufflinks. Therefore, we concluded that the isoform discovery performance of SLIDE is better than, or at least comparable to, that of Cufflinks.

The comparison results in the main paper are based on dataset 1 (Table 2.1). We did the same set of comparisons on datasets 2-4 (Table 2.1), and the results are summarized in Figures A.4 and A.5 (results on dataset 1 are in Figure 2.2B and C). From Figures A.4 and A.5, we observe that the comparison results on datasets 2-4 are consistent with the results on dataset 1.

A.3.2 Comparison between SLIDE and IsoLasso/NSMAP

Here, we compare SLIDE with two other isoform discovery methods with lasso-type sparse estimation: IsoLasso [27] and NSMAP [36].

SLIDE is different from IsoLasso [27] in three aspects. (i) IsoLasso enumerates isoforms based on a connectivity graph used by Scripture (10). This deterministic approach finds the

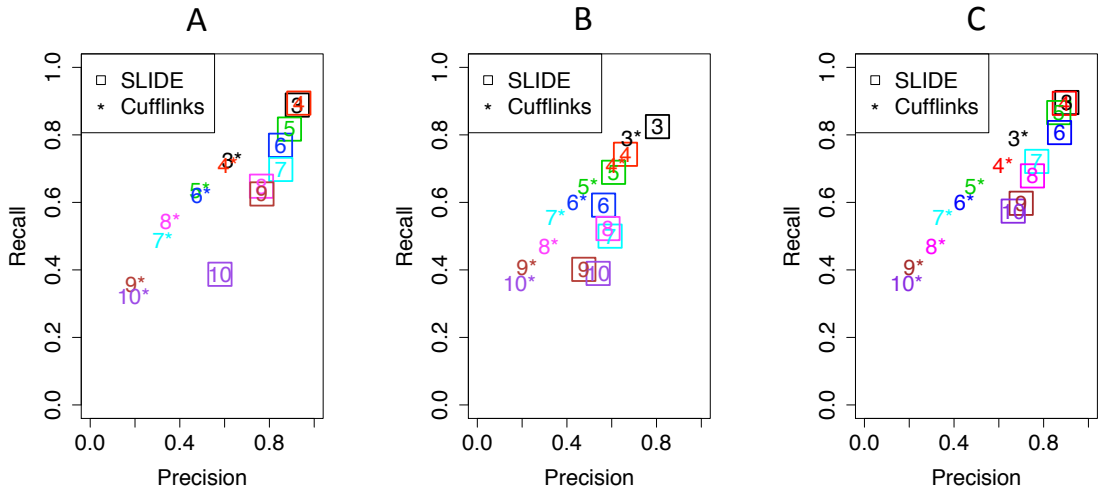

Figure A.4: Comparison of isoform discovery results by SLIDE (using genes and exons from the UCSC annotation) and Cufflinks. (A) Precision and recall rates of SLIDE and Cufflinks on dataset 2. The numbers in the figure are the group indices of genes (i.e., numbers of subexons). The squares and stars represent SLIDE and Cufflinks results, respectively. (B) Precision and recall rates of SLIDE and Cufflinks on dataset 3. The numbers, squares, and stars have the same meaning as in A. (C) Precision and recall rates of SLIDE and Cufflinks on dataset 4. The numbers, squares, and stars have the same meaning as in A.

longest paths indicated by connected paired-end reads and would not consider isoforms with alternative starts/ends (i.e., one isoform is a fragment of the other) as isoform candidates. (ii) IsoLasso uses a binary design matrix to relate reads to isoforms. It does not fully capture the quantitative relationship between read counts and isoform abundance. In contrast, our design matix uses conditional probabilities to relate read counts to isoform abundance, and is flexible in terms of incorporating different types of biological information into the modeling (e.g., using GC content to adjust nonuniform read coverage). (iii) IsoLasso performs isoform discovery and abundance estimation simultaneously with Lasso, a sparse estimation method. However, the penalization term in Lasso would introduce biases to the abundance estimates. To fix this issue, SLIDE uses a two-step approach that first discovers isoforms by sparse estimation and subsequently estimates the abundance of the discovered isoforms by nonnegative least squares that gives less biased estimates than Lasso does. (iv) Unlike IsoLasso, SLIDE favors isoforms with more exons in its sparse estimation. This is because we observe that RNA-Seq data noise often leads the linear model to fit with multiple isforms each with a small numbers of exons, contradicting with annotations. To counteract such data noise, we give less penalty to isoforms with more exons in the sparse estimation. In addition, IsoLasso builds isoform candidates from de novo exons directly assembled by mapped reads, without taking annotated gene and exon information into account.

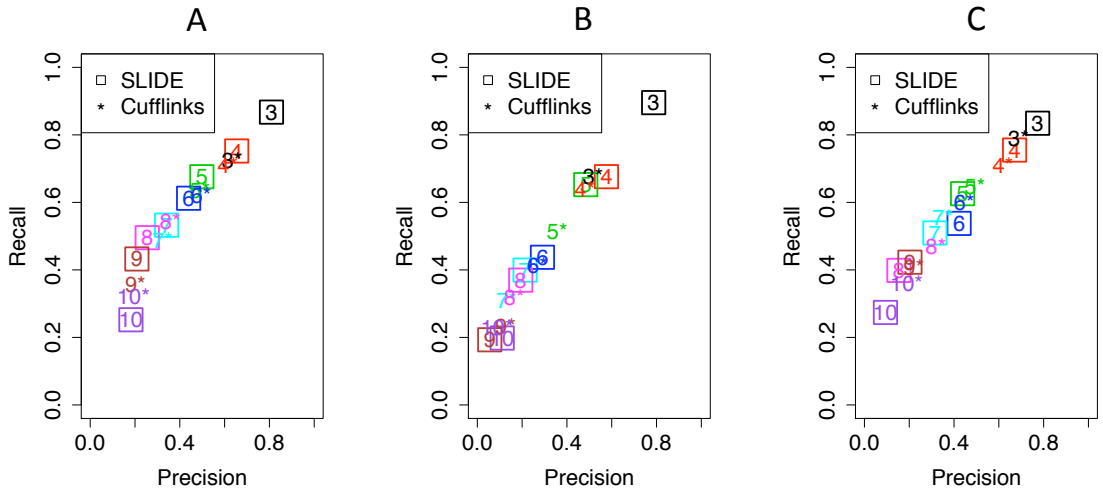

Figure A.5: Comparison of isoform discovery results by SLIDE (using de novo genes and exons assembled by Cufflinks) and Cufflinks. (A) Precision and recall rates of SLIDE and Cufflinks on dataset 2. The numbers in the figure are the group indices of genes (i.e., numbers of subexons). The squares and stars represent SLIDE and Cufflinks results, respectively. (B) Precision and recall rates of SLIDE and Cufflinks on dataset 3. The numbers, squares, and stars have the same meaning as in A. (C) Precision and recall rates of SLIDE and Cufflinks on dataset 4. The numbers, squares, and stars have the same meaning as in A.

We conducted three numerical comparisons between SLIDE and IsoLasso on isoform discovery based on the same data used in the comparison between SLIDE and Cufflinks. In the first comparison, we evaluated both methods in their default settings, where SLIDE builds isoforms from exons in UCSC annotations and IsoLasso finds isoforms from its de novo assembled exons. The discovered isoforms by either method are evaluated by UCSC annotations, where a discovered isoform is called to match an annotated isoform if they have the same number of exons and all of their exons overlap. Thus, this evaluation scheme is not sensitive to exon boundaries as long as a discovered isoform has exons in the same loci as exons of an annotated isoform. Precision and recall rates are calculated as described in the main paper. The comparison results in Figure A.6A show that SLIDE has better precision and recall rates than IsoLasso does. These results are similar to the first comparison results between SLIDE and Cufflinks in the main paper. The main reason is that both IsoLasso and Cufflinks find isoforms for de novo assembled genes, whose boundaries are sensitive to RNA-Seq data noise, especially to biases of junction read counts. These results suggest the importance of scrutinizing de novo assembled genes and exons with available annotations before performing isoform discovery; however, because SLIDE and IsoLasso do not start from the same set of genes and exons, this comparison is not a fair evaluation of their isoform assembly performance. Thus, we performed a second comparison based on isoforms discovered

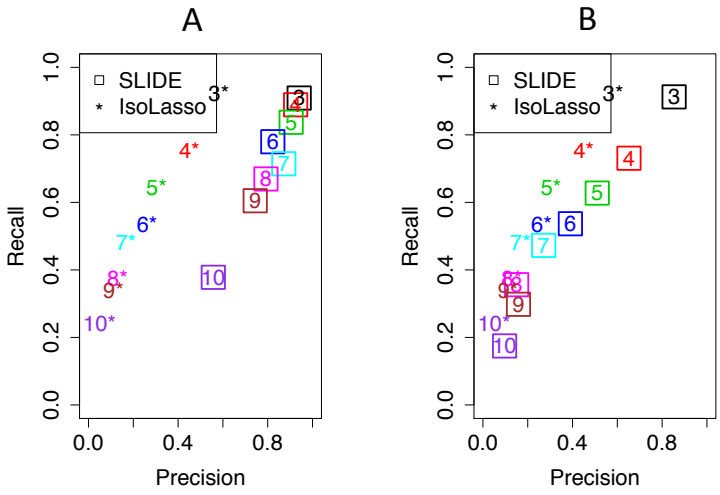

Figure A.6: Comparison of isoform discovery results by SLIDE and IsoLasso. (A) Precision and recall rates of SLIDE (using annotated genes/exons) and IsoLasso on dataset 1 (Table 2.1). The numbers in the figure are the group indices of genes (i.e., numbers of subexons). The squares and stars represent SLIDE and Cufflinks results, respectively. (B) Precision and recall rates of SLIDE (using IsoLasso assembled genes/exons) and IsoLasso on dataset 1 (Table 2.1). The numbers, squares, and stars have the same meaning as in A.

by either method from de novo genes and exons assembled by IsoLasso. Figure A.6B shows that SLIDE still has better precision rates than IsoLasso has for most genes. To further exclude the effects of disagreement between annotated and de novo assembled exons, we carried out a third comparison of the two methods using only the de novo assembled exons that agree with the annotation. We found that SLIDE has an average precision rate 0.85 and recall rate 0.91, whereas IsoLasso has an average precision rate 0.79 and recall rate 0.91. This again shows that SLIDE has higher precision than IsoLasso has in isoform assembly from the same set of de novo exons.

NSMAP is a Bayesian model-based method that estimates the abundance of isoform candidates as MAP (maximum a posteriori) estimates [36]. It is an extension of the maximum-likelihood abundance estimation method "statistical inferences for isoform expression in RNA-Seq" (SIIER) [24], in the sense of expanding parameters of interest from the abundance of annotated isoforms to that of all isoform candidates. NSMAP uses a Laplace prior to introduce sparseness and then discovers isoforms based on MAP estimates of the abundance of isoform candidates. NSMAP is similar to IsoLasso in four aspects. (i) Both methods use deterministic approaches to construct isoform candidates. NSMAP uses the minimal set of isoforms that can explain all junction reads, and it would miss isoforms with alternative starts/ends (i.e., one isoform is a fragment of the other) in its isoform candidate set. (ii) NSMAP is equivalent to IsoLasso in the optimization step. In the original Lasso paper, it

was suggested that Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace priors [34]. Unlike SLIDE, both NSMAP and IsoLasso do not favor isoforms with more exons in their sparse estimation, but they only keep the longest isoforms in their candidate sets. (iii) Both methods perform isoform discovery and abundance estimation simultaneously with sparse estimation. (iv) Both methods build isoforms from de novo assembled genes and exons. NSMAP constructs genes and exons de novo from read alignment output of Tophat. Therefore, the differences between SLIDE and NSMAP in methodology would be similar to those between SLIDE and IsoLasso. We tried to conduct numerical comparison between SLIDE and NSMAP. However, a code bug in the NSMAP package (Version 0.1.0) prohibited us from using it, and our attempts at contacting the authors were not successful.

A.4 More about mRNA Isoform Abundance Estimation on modENCODE Data

To evaluate the isoform abundance estimation accuracy of SLIDE without knowing the ground truth of isoform quantities in datasets 1-4 (Table 2.1), we compare SLIDE to two widely used methods: statistical inferences for isoform expression in RNA-Seq (SIIER) [24] and Cufflinks [16]. All three methods are used to estimate the isoform proportions of 317 chr3R genes with multiple isoforms in the UCSC annotation, and the total number of isoforms is 798. On dataset 1 (Table 2.1), the SLIDE and SIIER estimates have a correlation $R = 0.75$, and there are 25 genes with significantly inconsistent estimates between the two methods; i.e., $\hat{p}_{SLIDE} < 0.1$ and $\hat{p}_{SIIER} > 0.5$ or $\hat{p}_{SLIDE} > 0.5$ and $\hat{p}_{SIIER} < 0.1$. By detailed manual inspection, we find that among the 25 genes there are 20 genes whose SLIDE estimates agree better with the *paired-end bin* counts. For example, gene *CG9801* has five subexons with RPKMs 8.25, 5.57, 0, 3.92, and 3.16, respectively, and observed junctions between subexons 1-2, 2-4, and 4-5 in dataset 1. There are three isoforms (1,2), (1,2,5), and (1,2,4,5) of *CG9801* in the annotation. SLIDE estimates their proportions as 0.76, 0, and 0.24, respectively, whereas SIIERs estimates are 0, 0.55, and 0.45, respectively. Because there is no observed junction between subexons 2 and 5 and the expression levels of subexons 1 and 2 are higher than those of subexons 4 and 5, the SLIDE estimates seem more consistent with the data. The rest of the 25 genes include one gene whose SIIER estimates agree better with the paired-end bin counts, and four genes with ambiguous bin counts that cannot differentiate the two sets of estimates. An example of the ambiguous cases is gene *D1* with five subexons. In dataset 1, the RPKMs of the five subexons are 327.79, 326.01, 16.6, 6.23, and 0, respectively, and there are observed junction reads between subexons 1-2, 2-3, and 3-4. SLIDE estimates the proportions of annotated isoforms (1,2,3,4) and (1,2,3) as 0.02 and 0.98, respectively, whereas SIIER returns estimates 1 and 0, respectively. Based on the annotation, we would expect subexons 1, 2, and 3 to have similar expression levels; however, the observed expression in subexon 3 is significantly lower than that of subexons 1 and 2. So

the data seriously contradicts with the annotation. Hence, both SLIDE and SIIER cannot reasonably fit the data based on the annotation. After removing those 25 genes, we have a correlation $R = 0.88$ between the SLIDE and SIIER estimates.

In the comparison between SLIDE and Cufflinks, the correlation between their estimates on the proportions of the 798 isoforms is $R = 0.67$ on dataset 1. Similarly, we find 35 genes with significantly inconsistent estimates between SLIDE and Cufflinks, $\hat{p}_{SLIDE} < 0.1$ and $\hat{p}_{Cufflinks} > 0.5$ or $\hat{p}_{SLIDE} > 0.5$ and $\hat{p}_{Cufflinks} < 0.1$. Again by detailed manual inspection, we observe that 30 of them have SLIDE estimates that agree better with the paired-end bin counts, 3 have Cufflinks estimates that agree with the bin counts, and 2 have ambiguous bin counts such that both estimates are reasonable. After removing those 35 genes, we have a correlation $R = 0.85$ between the SLIDE and Cufflinks estimates.

A.5 More about the Exploration of Read/Fragment Length Effect

In the exploration of whether different read lengths would affect the isoform discovery results of SLIDE, we applied SLIDE to datasets 2 and 3 (Table 2.1), which are from the same Kc167 sample, with similar sequencing depth, but of read lengths 37 and 76 bp, respectively. Surprisingly, the precision and recall rates on the 37-bp data are higher than those on the 76-bp data. In the search for a possible explanation, we observed that the cDNA fragments in single-exon genes have different fragment length distributions in the two datasets: $N(166, 26^2)$ and $N(127, 13^2)$ for the 37-bp and 76-bp data, respectively.

To explore whether the read length or the fragment length has larger effects on the isoform discovery, we did a simulation study with two different read lengths (37 and 76 bp) and three different fragment length ranges (50-100 bp, 100-150 bp, and 150-200 bp). In each of the 50 simulation runs, 500 paired-end RNA-Seq reads are generated in each setting for each read length and each fragment length range. We applied SLIDE to the simulated data and summarized the precision and recall rates of each setting in Figure A.7. The figure illustrates that the increase in fragment lengths from 50-100 bp to 100-150 bp significantly improves the precision and recall rates of isoform discovery. Changes in fragment lengths from 100-150 bp to 150-200 bp also improve the precision and recall rates by increasing their means to some extent and decreasing the width of their confidence intervals. Compared to the fragment length changes, read length changes from 37 bp to 76 bp have smaller effects on the isoform discovery results.

A.6 Read Coverage vs. GC Content

It has been reported by several groups that read coverage has a strong correlation with GC content in high-throughput DNA sequencing data [110, 17]. As high-throughput sequencing technologies (e.g., DNA sequencing, RNA-Seq, ChIP-Seq, etc.) have similar characteristics

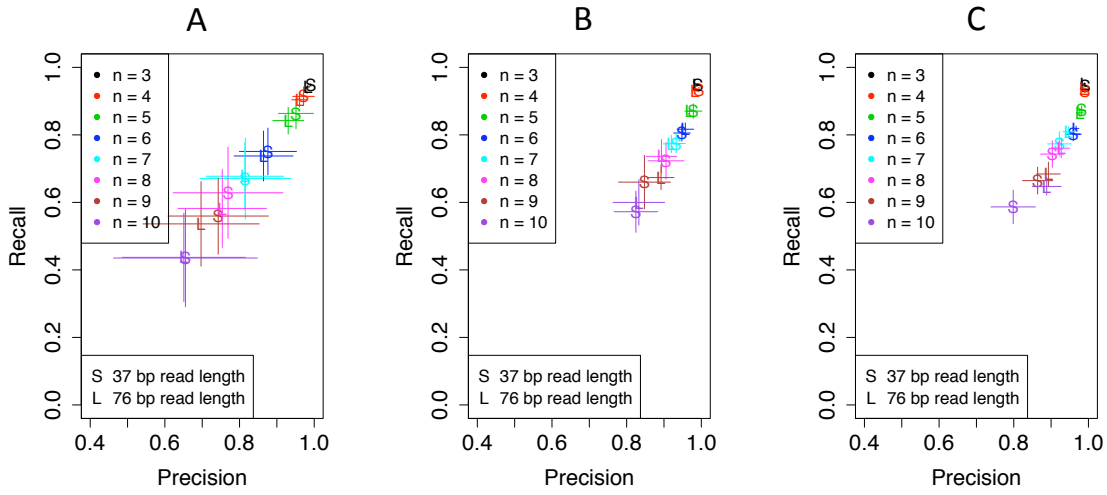

Figure A.7: Simulation study of read/fragment length effects on isoform discovery. (A) Precision and recall rates of SLIDE on simulated paired-end RNA-Seq data with fragment lengths in the range of 50-100 bp and two different read lengths (37 bp vs. 76 bp). 95% confidence intervals of precision and recall rates are shown as error bars parallel to the x and y axes, respectively. (B) Precision and recall rates of SLIDE on simulated paired-end RNA-Seq data with fragment lengths in the range of 100150 bp and two different read lengths (37 bp vs. 76 bp). The confidence intervals are shown in the same way as in A. (C) Precision and recall rates of SLIDE on simulated paired-end RNA-Seq data with fragment lengths in the range of 150-200 bp and two different read lengths (37 bp vs. 76 bp). The confidence intervals are shown in the same way as in A.

in the se- quencing step, many researchers believe that a strong correlation between read coverage and GC content also exists in RNA-Seq data [18, 19]. However, unlike DNA sequencing data, RNA-Seq read coverage varies in different transcribed regions and is mainly determined by expression levels and alternative splicing patterns of the regions [113]. It would be difficult to compare read coverage across subexons, which may occur in different transcripts and thus have different expression levels. To check the validity of using GC content correction in our SLIDE model, we study the relationship between RNA-Seq read coverage and GC content within subexons, using RNA-Seq reads on chr3R in dataset 1 (see Table 2.1). We use three different window sizes: 10 bp, 30 bp, and 50 bp. For every subexon, we calculate the correlation coefficient of its windowed average read coverage vs. GC content. Then, we calculate the percentage of subexons giving positive correlations among all the subexons with more than n windows ($n = 3, 10, \ldots, 100$), and find that the percentage increases as n increases. This trend is observed with all the three window sizes. The percentages for 10-bp windows are summarized in Table A.2. A histogram of the correlations in subexons with more than 100 windows is in Figure A.8. Because we expect that correlations calculated in

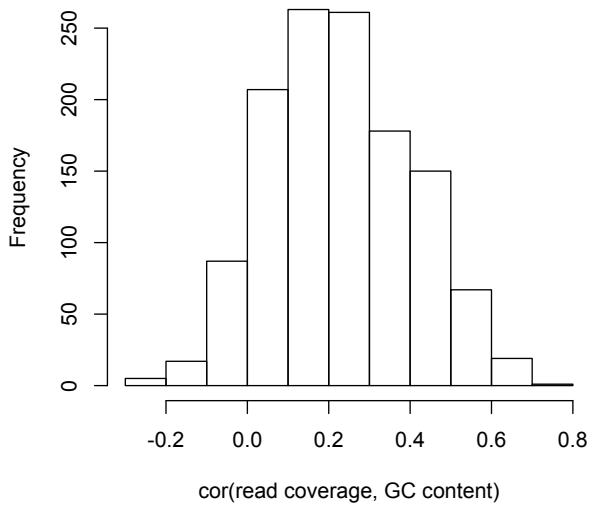

Figure A.8: A histogram of correlations between windowed read coverage and GC content in subexons containing more than 100 windows of 10-bp size.

n	3	10	20	40	60	80	100
Percentage	77.3%	78.6%	83.0%	87.9%	90.1%	91.0%	91.3%
Mean(R)	0.171	0.174	0.193	0.219	0.227	0.232	0.229

Table A.2: Percentages of subexons ($> n$ 10-bp windows) with positive correlation (R) between read coverage and GC content

subexons with more windows can better represent the relationship between read coverage and GC content, we conclude that there is a positive correlation between read coverage and GC content.

A.7 Some Other Detail about the Analysis

- In the simulation study of the main paper, we simulated reads from the 1,972 genes of 3-10 subexons (defined in Figure 2.1) on chr3R from *D. melanogaster* annotation (September 2010) of UCSC Genome Browser [29]. For each gene, reads are generated from the annotated isoforms, whose proportions p_k are randomly sampled from

$\{0, 0.1, \cdots, 0.9, 1\}$ subject to the constraint that $\sum_k p_k = 1$. For every gene, we simulate 500 reads in each run, with 50 runs in total.

- In the sparse estimation, we selected an optimal λ for each group of genes with n subexons ($n = 3, \cdots, 10$) by a stability criterion [37]. However, there are a small number of genes where zero isoforms were identified under the selected λ. For those genes, we reselected a gene-specific λ. In more details, we replace the previous λ by $\lambda^* = \max(\lambda - 0.1, \lambda/2)$ until non-trivial results were obtained.

- For isoform discovery, SLIDE uses sparse linear model estimation to find isoforms. We note that the linear model for paired-end data (Equation 2.3) is identifiable; i.e. $\mathbf{F}^T\mathbf{F}$ is invertible [111], for a few genes with 3-10 subexons of *D. melanogaster* (Table A.1). In those cases, we additionally attempted to use nonnegative least squares (NNLS), whose estimation results should be less biased than those of L_1 penalized estimation. However, compared to the penalized estimation results in the main paper, we found that the NNLS results include a lot of short isoforms as truncated fragments of isoforms in the UCSC annotation. In the example of gene *jumu*, whose three subexons have RPKMs 20.86, 42.62, and 25.97, respectively, there are observed junctions between subexons 1-2 and 2-3. NNLS discovered isoforms (1,2), (2,3), and (1,2,3) for *jumu*, whereas SLIDE only found the longest isoform (1,2,3), which agrees with the annotation. The possible reason of NNLS finding short isoforms is that RNA-Seq data have unexpected read coverage variation among exons in the same transcript [17, 18, 19].

Bibliography

[1] V.E. Velculescu et al. "Serial analysis of gene expression". In: *Science* 270.5235 (1995), p. 484.

[2] M. Schena et al. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". In: *Science* 270.5235 (1995), p. 467.

[3] S.C. Schuster. "Next-generation sequencing transforms today's biology". In: *Nature* 200.8 (2008).

[4] Signe Olivarius, Charles Plessy, and Piero Carninci. "High-throughput verification of transcriptional starting sites by Deep-RACE." In: *BioTechniques* 46.2 (2009), p. 130.

[5] Marcus Bantscheff et al. "Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present". In: *Analytical and bioanalytical chemistry* 404.4 (2012), pp. 939–965.

[6] Z. Wang, M. Gerstein, and M. Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1 (2009), pp. 57–63.

[7] J.A. Martin and Z. Wang. "Next-generation transcriptome assembly". In: *Nature Reviews Genetics* 12.10 (2011), pp. 671–682.

[8] S. MacArthur, X.Y. Li, J. Li, et al. "Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions". In: *Genome Biol* 10.7 (2009), R80.

[9] William W Fisher et al. "DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila". In: *Proceedings of the National Academy of Sciences* 109.52 (2012), pp. 21330–21335.

[10] J.J. Li et al. "Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation". In: *Proceedings of the National Academy of Sciences* 108.50 (2011), pp. 19867–19872.

[11] Björn Schwanhäusser et al. "Global quantification of mammalian gene expression control". In: *Nature* 473.7347 (2011), pp. 337–342.

[12] Jingyi Jessica Li, Peter J Bickel, and Mark D Biggin. "System Wide Analyses have Underestimated Protein Abundances and Transcriptional Importance in Animals". In: *arXiv preprint arXiv:1212.0587* (2012).

[13] Susan E Celniker et al. "Unlocking the secrets of the genome". In: *Nature* 459.7249 (2009), pp. 927–930.

[14] Sushmita Roy et al. "Identification of functional elements and regulatory circuits by Drosophila modENCODE". In: *Science* 330.6012 (2010), pp. 1787–1797.

[15] Mark B Gerstein et al. "Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project". In: *Science* 330.6012 (2010), pp. 1775–1787.

[16] Cole Trapnell et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature biotechnology* 28.5 (2010), pp. 511–515.

[17] Juliane C Dohm et al. "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing". In: *Nucleic acids research* 36.16 (2008), e105–e105.

[18] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. "Biases in Illumina transcriptome sequencing caused by random hexamer priming". In: *Nucleic acids research* 38.12 (2010), e131–e131.

[19] Jun Li, Hui Jiang, and Wing Hung Wong. "Method Modeling non-uniformity in short-read rates in RNA-Seq data". In: *Genome Biol* 11.5 (2010), R25.

[20] Adam Roberts et al. "Improving RNA-Seq expression estimates by correcting for fragment bias". In: *Genome Biol* 12.3 (2011), R22.

[21] Soohyun Lee et al. "Accurate quantification of transcriptome from RNA-Seq data by effective length normalization". In: *Nucleic acids research* 39.2 (2011), e9–e9.

[22] Ali Mortazavi et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature methods* 5.7 (2008), pp. 621–628.

[23] Jianxing Feng, Wei Li, and Tao Jiang. "Inference of isoforms from short sequence reads". In: *Research in Computational Molecular Biology*. Springer. 2010, pp. 138–157.

[24] Hui Jiang and Wing Hung Wong. "Statistical inferences for isoform expression in RNA-Seq". In: *Bioinformatics* 25.8 (2009), pp. 1026–1032.

[25] Bo Li et al. "RNA-Seq gene expression estimation with read mapping uncertainty". In: *Bioinformatics* 26.4 (2010), pp. 493–500.

[26] Hugues Richard et al. "Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments". In: *Nucleic acids research* 38.10 (2010), e112–e112.

[27] Wei Li, Jianxing Feng, and Tao Jiang. "Isolasso: a lasso regression approach to RNA-seq based transcriptome assembly". In: *Journal of Computational Biology* 18.11 (2011), pp. 1693–1707.

[28] P Flicek et al. "Ensembl 2011". In: *Nucleic Acids Res* 39.Suppl 1 (2011), pp. D800–D806.

[29] Pauline A Fujita et al. "The UCSC genome browser database: update 2011". In: *Nucleic acids research* 39.suppl 1 (2011), pp. D876–D882.

[30] Mitchell Guttman et al. "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs". In: *Nature biotechnology* 28.5 (2010), pp. 503–510.

[31] Michael A Frohman, Michael K Dush, and Gail R Martin. "Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer". In: *Proceedings of the National Academy of Sciences* 85.23 (1988), pp. 8998–9002.

[32] Toshiyuki Shiraki et al. "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage". In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15776–15781.

[33] Mark D Adams et al. "Complementary DNA sequencing: expressed sequence tags and human genome project". In: *Science* 252.5013 (1991), pp. 1651–1656.

[34] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[35] Song Liu et al. "A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species". In: *Nucleic Acids Research* 39.2 (2011), pp. 578–588.

[36] Zheng Xia et al. "NSMAP: A method for spliced isoforms identification and quantification from RNA-Seq". In: *BMC bioinformatics* 12.1 (2011), p. 162.

[37] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010), pp. 417–473.

[38] Corinne Dahinden et al. "Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries". In: *BMC bioinformatics* 8.1 (2007), p. 476.

[39] J Goeman. "penalized. L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model". In: *R Package version 0.9-31* (2010). Available at `http://cran.r-project.org/web/packages/penalized/`.

[40] Katharine M Mullen and Ivo HM van Stokkum. "nnls: The Lawson-Hanson Algorithm for Non-Negative Least Squares (NNLS)". In: *R Package version 1.3* (2010). Available at `http://cran.r-project.org/web/packages/nnls/`.

[41] Edward H Kislauskis, Xiao-chun Zhu, and Robert H Singer. "β-Actin messenger RNA localization and protein synthesis augment cell motility". In: *The Journal of cell biology* 136.6 (1997), pp. 1263–1270.

[42] Mark D Biggin. "Animal transcription networks as highly connected, quantitative continua". In: *Developmental cell* 21.4 (2011), pp. 611–626.

[43] Nicholas T Ingolia, Liana F Lareau, and Jonathan S Weissman. "Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes". In: *Cell* 147.4 (2011), pp. 789–802.

[44] Raquel de Sousa Abreu et al. "Global signatures of protein and mRNA expression levels". In: *Molecular BioSystems* 5.12 (2009), pp. 1512–1526.

[45] Michal Rabani et al. "Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells". In: *Nature biotechnology* 29.5 (2011), pp. 436–442.

[46] Matthias W Hentze and Lukas C Kühn. "Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress". In: *Proceedings of the National Academy of Sciences* 93.16 (1996), pp. 8175–8182.

[47] Edward Yang et al. "Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes". In: *Genome research* 13.8 (2003), pp. 1863–1872.

[48] Andrew C Hsieh et al. "The translational landscape of mTOR signalling steers cancer initiation and metastasis". In: *Nature* 485.7396 (2012), pp. 55–61.

[49] Oliver Hobert. "Gene regulation by transcription factors and microRNAs". In: *Science Signaling* 319.5871 (2008), p. 1785.

[50] Ramesh S Pillai, Suvendra N Bhattacharyya, and Witold Filipowicz. "Repression of protein synthesis by miRNAs: how many mechanisms?" In: *Trends in cell biology* 17.3 (2007), pp. 118–126.

[51] Vincenzo Alessandro Gennarino et al. "Identification of microRNA-regulated gene networks by expression analysis of target genes". In: *Genome research* 22.6 (2012), pp. 1163–1172.

[52] Matthias Selbach et al. "Widespread changes in protein synthesis induced by microRNAs". In: *Nature* 455.7209 (2008), pp. 58–63.

[53] Daehyun Baek et al. "The impact of microRNAs on protein output". In: *Nature* 455.7209 (2008), pp. 64–71.

[54] Christine Vogel et al. "Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line". In: *Molecular systems biology* 6.1 (2010).

[55] Martin Beck et al. "The quantitative proteome of a human cell line". In: *Molecular systems biology* 7.1 (2011).

[56] Christine Vogel and Edward M Marcotte. "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses". In: *Nature Reviews Genetics* 13.4 (2012), pp. 227–232.

[57] Tobias Maier, Marc Güell, and Luis Serrano. "Correlation of mRNA and protein in complex biological samples". In: *FEBS letters* 583.24 (2009), pp. 3966–3973.

[58] AKEMI Hanamura et al. “Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors.” In: *Rna* 4.4 (1998), pp. 430–444.

[59] Simon G Gregory et al. “A physical map of the mouse genome”. In: *Nature* 418.6899 (2002), pp. 743–750.

[60] Alan Wolffe. *Chromatin: structure and function*. Academic press, 1998.

[61] Michael F Princiotta et al. “Quantitating protein synthesis, degradation, and endogenous antigen processing”. In: *Immunity* 18.3 (2003), pp. 343–354.

[62] R Brosi, HP Hauri, and A Krämer. “Separation of splicing factor SF3 into two components and purification of SF3a activity.” In: *Journal of Biological Chemistry* 268.23 (1993), pp. 17640–17646.

[63] Philip G Wong et al. “Cdc45 limits replicon usage from a low density of preRCs in mammalian cells”. In: *PLoS One* 6.3 (2011), e17533.

[64] Hiroshi Kimura et al. “Quantitation of RNA polymerase II and its transcription factors in an HeLa cell: little soluble holoenzyme but significant amounts of polymerases attached to the nuclear substructure”. In: *Molecular and cellular biology* 19.8 (1999), pp. 5383–5392.

[65] Marlis Zeiler et al. “A protein epitope signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines”. In: *Molecular & Cellular Proteomics* 11.3 (2012).

[66] Tilman Borggrefe et al. “Quantitation of the RNA polymerase II transcription machinery in yeast”. In: *Journal of Biological Chemistry* 276.50 (2001), pp. 47150–47153.

[67] Ron Milo et al. “BioNumbersthe database of key numbers in molecular and cell biology”. In: *Nucleic acids research* 38.suppl 1 (2010), pp. D750–D753.

[68] Norikazu Aoyagi and David A Wassarman. “Developmental and transcriptional consequences of mutations in Drosophila TAF II 60”. In: *Molecular and cellular biology* 21.20 (2001), pp. 6808–6819.

[69] Joel C Eissenberg et al. “dELL is an essential RNA polymerase II elongation factor with a general role in development”. In: *Proceedings of the National Academy of Sciences* 99.15 (2002), pp. 9894–9899.

[70] Adam M Deutschbauer et al. “Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast”. In: *Genetics* 169.4 (2005), pp. 1915–1925.

[71] Dong-Uk Kim et al. “Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe”. In: *Nature biotechnology* 28.6 (2010), pp. 617–623.

[72] Sina Ghaemmaghami et al. “Global analysis of protein expression in yeast”. In: *Nature* 425.6959 (2003), pp. 737–741.

[73] Peng Lu et al. “Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation”. In: *Nature biotechnology* 25.1 (2006), pp. 117–124.

[74] Juliane C Dohm et al. “Substantial biases in ultra-short read data sets from high-throughput DNA sequencing”. In: *Nucleic acids research* 36.16 (2008), e105–e105.

[75] Ming-Sin Cheung et al. “Systematic bias in high-throughput sequencing data and its correction by BEADS”. In: *Nucleic acids research* 39.15 (2011), e103–e103.

[76] Gary K Geiss et al. “Direct multiplexed measurement of gene expression with color-coded probe pairs”. In: *Nature biotechnology* 26.3 (2008), pp. 317–325.

[77] Srilatha Kuntumalla et al. “Comparison of two label-free global quantitation methods, APEX and 2D gel electrophoresis, applied to the Shigella dysenteriae proteome”. In: *Proteome science* 7.1 (2009), p. 22.

[78] Daniel Hebenstreit et al. “RNA sequencing reveals two major classes of gene expression levels in metazoan cells”. In: *Molecular systems biology* 7.1 (2011).

[79] Daniel Hebenstreit et al. “Duel of the fates: the role of transcriptional circuits and noise in CD4+ cells”. In: *Current Opinion in Cell Biology* (2012).

[80] Jacek R Wi&sacute, Kamila Du&sacute Ostasiewicz, Dorota F Zieli&nacute, et al. “Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma”. In: *Molecular systems biology* 8.1 (2012).

[81] Peter J Bickel and Kjell A Doksum. *Mathematical Statistics, volume I*. Prentice Hall Englewood Cliffs, NJ, 2001.

[82] Ingram Olkin and Jeremy Finn. “Testing correlated correlations”. In: *Psychological Bulletin* 108.2 (1990), pp. 330–333.

[83] ENCODE Project Consortium et al. “A users guide to the encyclopedia of DNA elements (ENCODE)”. In: *PLoS Biol* 9.4 (2011), e1001046.

[84] *Caenorhabditis elegans* Sequencing Consortium et al. “Genome sequence of the nematode *C. elegans*: a platform for investigating biology”. In: *Science* 282.2012 (1998), p. 2018.

[85] Mark D Adams et al. “The genome sequence of *Drosophila melanogaster*”. In: *Science* 287.5461 (2000), pp. 2185–2195.

[86] Michelle N Arbeitman et al. “Gene expression during the life cycle of *Drosophila melanogaster*”. In: *Science* 297.5590 (2002), pp. 2270–2275.

[87] Min Jiang et al. “Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*”. In: *Proceedings of the National Academy of Sciences* 98.1 (2001), pp. 218–223.

[88] Alex T Kalinka et al. “Gene expression divergence recapitulates the developmental hourglass model”. In: *Nature* 468.7325 (2010), pp. 811–814.

[89] Stuart K Kim et al. "A gene expression map for *Caenorhabditis elegans*". In: *Science Signaling* 293.5537 (2001), p. 2087.

[90] Viktor Stolc et al. "A gene expression map for the euchromatic genome of *Drosophila melanogaster*". In: *Science* 306.5696 (2004), pp. 655–660.

[91] Alex T Kalinka and Pavel Tomancak. "The evolution of early animal embryos: conservation or divergence?" In: *Trends in ecology & evolution* (2012).

[92] Toshie Kai and Allan Spradling. "Differentiating germ cells can revert into functional stem cells in *Drosophila melanogaster* ovaries". In: *Nature* 428.6982 (2004), pp. 564–569.

[93] Jose F de Celis, Marta Llimargas, and Jordi Casanova. "Ventral veinless, the gene encoding the Cf1a transcription factor, links positional information and cell differentiation during embryonic and imaginal development in *Drosophila melanogaster*". In: *Development* 121.10 (1995), pp. 3405–3416.

[94] Alain Nepveu. "Role of the multifunctional CDP/Cut/Cux homeodomain transcription factor in regulating differentiation, cell growth and development." In: *Gene* 270.1-2 (2001), p. 1.

[95] Jeffrey C Way and Martin Chalfie. "*mec-3*, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in *C. elegans*". In: *Cell* 54.1 (1988), pp. 5–16.

[96] Michael Finney, Gary Ruvkun, and H Robert Horvitz. "The *C. elegans* cell lineage and differentiation gene *unc-86* encodes a protein with a homeodomain and extended similarity to transcription factors". In: *Cell* 55.5 (1988), pp. 757–769.

[97] Michael F Lin et al. "Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes". In: *Genome research* 17.12 (2007), pp. 1823–1836.

[98] Paul Flicek et al. "Ensembl 2012". In: *Nucleic acids research* 40.D1 (2012), pp. D84–D90.

[99] Karen Yook et al. "WormBase 2012: more genomes, more data, new website". In: *Nucleic acids research* 40.D1 (2012), pp. D735–D741.

[100] Neal S Holter et al. "Fundamental patterns underlying gene expression profiles: simplicity from complexity". In: *Proceedings of the National Academy of Sciences* 97.15 (2000), pp. 8409–8414.

[101] Lucy Cherbas et al. "The transcriptional diversity of 25 Drosophila cell lines". In: *Genome research* 21.2 (2011), pp. 301–314.

[102] Susan E Lott et al. "Noncanonical compensation of zygotic X transcription in early *Drosophila melanogaster* development revealed through single-embryo RNA-seq". In: *PLoS biology* 9.2 (2011), e1000590.

[103] Sudhir Nayak, Johnathan Goree, and Tim Schedl. "*fog-2* and the evolution of self-fertile hermaphroditism in Caenorhabditis". In: *PLoS biology* 3.1 (2004), e6.

[104] Michael Ashburner et al. “Gene Ontology: tool for the unification of biology”. In: *Nature genetics* 25.1 (2000), pp. 25–29.

[105] Mikel E Aranguren et al. “Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL”. In: *BMC bioinformatics* 8.1 (2007), p. 57.

[106] Sergio Barberan-Soler and Alan M Zahler. “Alternative splicing regulation during *C. elegans* development: splicing factors as regulated targets”. In: *PLoS genetics* 4.2 (2008), e1000001.

[107] Nathan Salomonis et al. “Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation”. In: *Proceedings of the National Academy of Sciences* 107.23 (2010), pp. 10514–10519.

[108] Mark B Gerstein et al. “Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project”. In: *Science* 330.6012 (2010), pp. 1775–1787.

[109] Sushmita Roy et al. “Identification of functional elements and regulatory circuits by Drosophila modENCODE”. In: *Science* 330.6012 (2010), pp. 1787–1797.

[110] David James Hiller. “Alternative splicing analysis using RNA-seq data”. PhD thesis. Stanford University, 2010.

[111] David Hiller et al. “Identifiability of isoform deconvolution from junction arrays and RNA-Seq”. In: *Bioinformatics* 25.23 (2009), pp. 3056–3059.

[112] Sika Zheng and Liang Chen. “A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level”. In: *Nucleic Acids Research* 37.10 (2009), e75–e75.

[113] Sudeep Srivastava and Liang Chen. “A two-parameter generalized Poisson model to improve the analysis of RNA-seq data”. In: *Nucleic Acids Research* 38.17 (2010), e170–e170.